

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Bruna Roberta Seewald da Silva

UMA PROPOSTA DE CONCEITOS DE JUSTIÇA APLICADAS A
MODELOS DE ANÁLISE DE SOBREVIVÊNCIA

Belo Horizonte
2021

Bruna Roberta Seewald da Silva

**UMA PROPOSTA DE CONCEITOS DE JUSTIÇA APLICADAS A
MODELOS DE ANÁLISE DE SOBREVIVÊNCIA**

Versão final

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Flavio Vinicius Diniz de Figueiredo

Belo Horizonte
2021

Silva. Bruna Roberta Seewald da;

S586u Uma proposta de conceitos de justiça aplicadas a modelos de análise de sobrevivência [manuscrito] Bruna Roberta Seewald da Silva. /.— 2021.
117 f. il.

Orientador: Flavio Vinicius Diniz de Figueiredo
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação

Referências: f. 100-105.

1. Computação – Teses. 2. Análise de sobrevivência-Teses. 3. Justiça – Aprendizado de máquina – Teses. I Figueiredo, Flavio Vinicius Diniz.. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. IV. Título.

CDU 519.6*63(043)



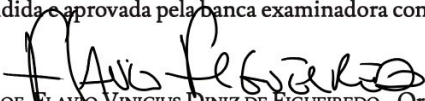
UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

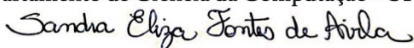
FOLHA DE APROVAÇÃO

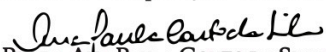
Uma proposta de conceitos de justiça aplicadas a modelos de análise de sobrevivência

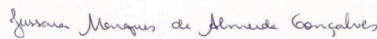
BRUNA ROBERTA SEEWALD DA SILVA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:


PROF. FLAVIO VINICIUS DINIZ DE FIGUEIREDO - Orientador
Departamento de Ciência da Computação - UFMG


PROFA. SANDRA ELIZA FONTES DE AVILA
Instituto de Computação - UNICAMP


PROFA. ANA PAULA COUTO DA SILVA
Departamento de Ciência da Computação - UFMG


PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 3 de Dezembro de 2021.

*À todas as pessoas que me inspiram a ser a melhor versão
de mim mesma.*

Agradecimentos

Essa dissertação é o resultado de uma longa jornada, que por muitas e muitas vezes se mostrou incerta e difícil. O caminho até aqui foi cansativo porém cheio de recompensas e nada disso seria possível sem a presença de todas as pessoas que nos querem bem.

Agradeço primeiramente a Luiza, minha eterna fonte de inspiração. Obrigada por me apresentar à UFMG e por sempre me incentivar a ir além, mesmo quando eu constantemente duvido das minhas capacidades. Sem você, nada disso teria acontecido.

Agradeço do fundo do meu coração a minha mãe, ao meu irmão e a minha avó. Vocês são a minha base fundamental. Sempre estiveram ao meu lado, em todos os momentos e em todas as curvas malucas da minha vida. O carinho e compreensão de vocês em toda essa jornada me deu forças para continuar. Esse título é nosso!

Ao meu querido pai, que esteja onde estiver, sempre me incentivou a estudar. Levo comigo as tuas palavras de incentivo e a certeza que o caminho do conhecimento é eterno e nunca deixarei de trilhá-lo.

Aos amigos que o DCC me deu, fontes infinitas de gargalhadas e momentos inesquecíveis. O caminho foi difícil mas com vocês certamente foi mais leve. Sou profundamente grata por ter vocês ao meu lado.

Ao meu orientador, Prof. Flavio, pela sensibilidade e compreensão durante tempos tão difíceis e por ter contribuído tanto para a minha formação.

E por fim à CAPES pela bolsa concedida.

Muito obrigada!

*“ Valeu a pena? Tudo vale a pena
Se a alma não é pequena.
Quem quer passar além do Bojador
Tem que passar além da dor.
Deus ao mar o perigo e o abismo deu,
Mas nele é que espelhou o céu.”*

(Fernando Pessoa em "Mar Português"(1934))

Resumo

O crescente uso de Aprendizado de Máquina (AM) no nosso cotidiano traz preocupações sobre os possíveis impactos que tais ferramentas podem causar na sociedade. Atualmente, já vemos algoritmos sendo utilizados para auxiliar no processo de contratação de pessoas e na identificação de possíveis criminosos através de imagens. Do ponto de vista socioeconômico, as decisões tomadas em situações como essas são cruciais. Portanto, é essencial que não haja um comportamento discriminatório em relação a certos grupos e/ou populações. Essa dissertação irá propor uma ponte entre justiça em AM sob uma nova perspectiva, de análise de sobrevivência (i.e., modelos de Cox e variações com Aprendizado Profundo). Para realizar nosso objetivo, fizemos uma revisão em ambas as literaturas. Depois, apresentamos quatro propostas de definição de justiça para análise de sobrevivência. A primeira e a segunda proposta, denominadas divergência em paridade demográfica, focaram na divergência entre as curvas empíricas e as curvas preditas. A terceira proposta, denominada discriminação causal, consistiu de verificar o erro a partir do cálculo do c-index, ao se alterar nos dados o grupo de interesse. E a última proposta é uma nova métrica, denominada justiça de filas, que compara indivíduos de grupos diferentes ao mesmo tempo. Em seguida aplicamos essas propostas nas três bases de dados: a hospitalar MIMIC, além de duas bases de dados reincidência criminal, a Rossi e COMPAS. Entre os resultados, na base MIMIC-III foram encontrados vieses nas proposta de divergência em paridade demográfica, divergência em paridade demográfica condicionada e justiça de filas. Um exemplo foi a divergência entre as curvas empíricas e preditas de pessoas negras com câncer, que também apareceu nos recortes para mulheres negras e homens negros com câncer. Nas bases Rossi e COMPAS os vieses apareceram junto a métrica de justiça de filas, que foi a única a identificar vieses em todas as bases analisadas. Além de trazer a discussão de justiça em três contextos diferentes, no momento, essa dissertação é o primeiro trabalho a trazer tal discussão para modelos de Cox.

Palavras-chave: Justiça em aprendizado de máquina, Análise de sobrevivência, modelos de Cox.

Abstract

The high demand for machine learning (ML) in our lives brings concern about the possible impacts that such tools may cause on society. Nowadays, algorithms are being used to assist hiring processes and identify criminals through images. From a societal standpoint, decisions made in these situations are critical. Therefore, it is essential that there is no discriminatory behavior against certain groups. To assess this issue, this work proposes to build a bridge between fairness in ML under a new perspective, from survival analysis (e.g., Cox models and variations with deep learning). To accomplish our goal, we have reviewed literature from both areas. Next, we introduced four proposal definitions of fairness for survival analysis. The first and second proposal are nominated divergence in demographic parity. Both of them focus in the difference between empirical and predicted curves. The third proposal, called casual discrimination, verifies the error of calculating the c-index when we change data for specifics group. The last proposal is a new metric, called "justiça de filas", which compares individuals from different groups at the same time. After that, we applied these proposals on three different databases: the first one was from the health domain, MIMIC-III, and the other two were from the criminal domain, Rossi and COMPAS. In MIMIC-III database, bias appeared in the divergence in demographic parity proposal, divergence in conditional demographic parity proposal and "justiça de filas". For example, there was a difference between the empirical and predicted curves for blacks with cancer. The same happened with black women and black men with cancer. In Rossi and COMPAS databases, situations with bias appeared using the "justiça de filas" metric, which was responsible for identified bias in all databases analyzed. In addition to the discussion in these three different contexts, at this moment, this dissertation is the first one to bring this to Cox models.

Keywords: Fairness in machine learning, Survival analysis, Cox models.

Lista de Figuras

1.1	Heatmap com trabalhos anteriores em justiça agrupadas por domínio. Fonte: Adaptado de Mehrabi et al. (2019)	17
2.1	Ilustração demonstrando um problema de análise de sobrevivência. Fonte: adaptado de Wang et al. (2019)	22
2.2	Ilustração apresentando relações entre as funções $S(t)$, $F(t)$ e $f(t)$. Fonte: adaptado de Wang et al. (2019)	24
2.3	Curva da banheira	25
2.4	Exemplo de uma curva Kaplan-Meier	26
3.1	Tabelas com dados do Compas. Fonte: Adaptado de (Gummadi, 2019)	33
4.1	Divergência em paridade demográfica para modelos de análise de sobrevivência	43
4.2	Divergência em paridade demográfica condicionada para modelos de análise de sobrevivência	44
4.3	Cálculo da métrica justiça de filas	45
5.1	Sumário estatístico para a população de pacientes selecionados	52
5.2	Gráficos de densidade para gênero e seguro de saúde	52
5.3	Gráfico com distribuição dos tempos de hospitalização	53
5.4	Pontuação Oasis para gênero e raça	53
5.5	Sumário estatístico para a base Rossi	55
5.6	Gráfico de censura para 25 indivíduos da base Rossi.	55
5.7	Análise das distribuições de idade e prisões anteriores	56
5.8	Sumário estatístico para a base COMPAS	56
5.9	Gráfico de censura para 25 indivíduos da base COMPAS.	57
5.10	Análise das distribuições do atributo <i>race</i> para base COMPAS	57
5.11	Gráfico do erro no treinamento e teste em função da complexidade do modelo. Adaptado de Friedman et al. (2001)	59
5.12	Validação cruzada. Adaptado de Pedregosa et al. (2011)	60

5.13	Diagrama que mostra o fluxo da modelagem usando validação cruzada e otimização de parâmetros. Adaptado de Pedregosa et al. (2011)	60
5.14	Boxplot comparativo entre os algoritmos	65
5.15	Estatística sobre o resultado do modelo de Cox na base Rossi	66
5.16	Impacto das variáveis na curva de sobrevivência	67
5.17	Cenário 1 - Estatística sobre o resultado do modelo de Cox na base COMPAS	68
5.18	Cenário 2 - Estatística sobre o resultado do modelo de Cox na base COMPAS	69
6.1	Curvas de sobrevivência divididas em gênero e raça	71
6.2	Curvas de sobrevivência para gênero/raça	72
6.3	Curva predita para gênero e raça	73
6.4	Curva predita para gênero-raça	73
6.5	Curva de sobrevivência com o estimador Kaplan-Meier	75
6.6	Paridade demográfica	76
6.7	Paridade demográfica considerando raça	76
6.8	Curva de sobrevivência com o estimador Kaplan-Meier para a base COMPAS	78
6.9	Curva predita considerando raça para a base COMPAS	78
6.10	Curvas KM para a proposta 2 da base MIMIC-III	80
6.11	Curva predita para a proposta 2 da base MIMIC-III	82
6.12	Curvas KM para a definição de divergência em paridade demográfica condicionada na base Rossi	84
6.13	Paridade demográfica condicionada na base Rossi	86
6.14	Curva KM para proposta 2 na base COMPAS	88
6.15	Curvas preditas para a proposta 2 na base COMPAS	89
A.1	Curvas KM com base no gênero para a proposta 2 na base MIMIC-III . . .	108
A.2	Curvas KM com base na raça para a proposta 2 na base MIMIC-III	109
A.3	Curvas KM com base na raça e gênero feminino para a proposta 2 na base MIMIC-III	110
A.4	Curvas KM com base na raça e gênero masculino para a proposta 2 na base MIMIC-III	111
A.5	Curvas preditas com base no gênero para a proposta 2 na base MIMIC-III	113
A.6	Curvas preditas com base na raça para a proposta 2 na base MIMIC-III . .	114
A.7	Curvas preditas com base na raça e gênero feminino para a proposta 2 na base MIMIC-III	115
A.8	Curvas preditas com base na raça e gênero masculino para a proposta 2 na base MIMIC-III	116

Lista de Tabelas

2.1	Notações em análise de sobrevivência	21
2.2	Funções densidade, sobrevivência e de risco para distribuições que usam métodos paramétricos	27
3.1	Visão geral das definições de justiça	36
3.2	Notação para classificação binária	37
3.3	Visão geral dos projetos endereçando o problema de justiça em AM	40
5.1	Atributos presentes na tabela final usada para os experimentos	51
5.2	Atributos presentes na tabela final usada para os experimentos	54
5.3	Espaço de busca para os hiperparâmetros	61
5.4	Espaço de busca para os hiperparâmetros	61
5.5	Espaço de busca para os hiperparâmetros	62
5.6	Espaço de busca para os hiperparâmetros	62
5.7	Hiperparâmetros adicionais no DeepHit	64
6.1	Tabela com resultados do teste LogRank para a proposta 1 da base MIMIC-III	72
6.2	Resultados do teste KS para a proposta 1 da base MIMIC-III	73
6.3	Resultados para a proposta de divergência em paridade demográfica	74
6.4	Resultados para a proposta de divergência em paridade demográfica para a base Rossi	75
6.5	Tabela com resultados do teste LogRank para a base Rossi	75
6.6	Resultados do teste KS para proposta 2 da base Rossi	77
6.7	Resultados para a proposta de divergência em paridade demográfica na base COMPAS	77
6.8	Tabela com resultados do teste LogRank para a proposta 1 da base COMPAS	77
6.9	Tabela com resultados do teste KS para a proposta 1 da base COMPAS	77
6.10	Tabela com resultados do teste LogRank para a proposta 2 da base MIMIC-III	81
6.11	Resultados do teste KS para proposta 2 da base MIMIC-III	81

6.12	Resultados para a proposta de divergência de paridade demográfica condicionada na base MIMIC-III	83
6.13	Tabela com resultados do teste LogRank para a proposta 2 da base Rossi .	85
6.14	Resultados do teste KS para proposta de divergência em paridade demográfica condicionada na base Rossi	86
6.15	Resultados para a proposta de divergência em paridade demográfica condicionada na base Rossi	87
6.16	Resultados do teste LogRank para proposta 2 na base COMPAS	88
6.17	Resultados do teste KS para proposta 2 na base COMPAS	89
6.18	Resultados para a proposta de divergência em paridade demográfica condicionada na base COMPAS	90
6.19	Resultados para a proposta de discriminação causal	91
6.20	Resultados para a proposta de discriminação causal para a base Rossi . . .	91
6.21	Resultados para a proposta de discriminação causal para a base COMPAS	92
6.22	Resultados para a quarta proposta	93
6.23	Resultados para a quarta proposta junto a base Rossi	94
6.24	Resultados para a quarta proposta junto a base COMPAS	95
A.1	Tabela com resultados do teste LogRank para a proposta 2 da base MIMIC-III	112
A.2	Resultados do teste KS para proposta 2 da base MIMIC-III	117

Sumário

1	Introdução	16
2	Análise de sobrevivência	20
2.1	Definições	21
2.1.1	Censura	21
2.1.2	Função de sobrevivência	23
2.1.3	Função de risco	24
2.2	Métodos estatísticos tradicionais	25
2.2.1	Métodos não-paramétricos	25
2.2.2	Métodos paramétricos	27
2.2.3	Métodos semi-paramétricos	28
2.2.4	Cox de riscos proporcionais	28
2.3	Métricas de Avaliação	29
2.3.1	C-Index	29
2.3.2	IPCW Brier Score	30
3	Justiça em aprendizado de máquina	32
3.1	Visão geral sobre aprendizado de máquina e justiça	32
3.1.1	Visão geral sobre aprendizado supervisionado	32
3.1.2	Exemplo de uma aplicação real	33
3.2	Justiça	34
3.2.1	Discriminação por tratamento desigual e discriminação por im- pacto desigual	35
3.3	Principais definições de justiça	36
3.4	Soluções para mitigar o problema de justiça	39
4	Definições de justiça para modelos de análise de sobrevivência	41
4.1	Propostas	41

4.1.1	Proposta 1: Divergência em paridade demográfica / estatística .	42
4.1.2	Proposta 2: Divergência em paridade demográfica condicionada	43
4.1.3	Proposta 3: Discriminação causal	43
4.1.4	Proposta 4: Justiça de filas	44
5	Experimentos e resultados	47
5.1	Bases de dados	47
5.2	Análises exploratórias	49
5.2.1	MIMIC-III	49
5.2.2	Rossi	54
5.2.3	COMPAS	56
5.3	Modelos	58
5.3.1	MIMIC-III	58
5.3.2	Rossi	64
5.3.3	COMPAS	64
5.4	Resultados	64
5.4.1	MIMIC-III	65
5.4.2	Rossi	66
5.4.3	COMPAS	68
6	Discussões sobre justiça	70
6.1	Proposta 1: Divergência em paridade demográfica	70
6.1.1	MIMIC-III	70
6.1.2	Rossi	74
6.1.3	COMPAS	77
6.2	Proposta 2: Divergência em paridade demográfica condicionada	79
6.2.1	MIMIC-III	79
6.2.2	Rossi	84
6.2.3	COMPAS	87
6.3	Proposta 3: Discriminação causal	90
6.3.1	MIMIC-III	90
6.3.2	Rossi	91
6.3.3	COMPAS	92
6.4	Proposta 4: Justiça de filas	93
6.4.1	MIMIC-III	93
6.4.2	Rossi	94
6.4.3	COMPAS	95

7 Conclusão	97
Referências Bibliográficas	100
Apêndice A Resultados complementares	106

Capítulo 1

Introdução

Quantas decisões você toma por dia ? São inúmeras as decisões que tomamos todos os dias, desde as mais pequenas e simples às mais complexas. No entanto, com o passar do tempo, nós criamos alternativas para automatizar várias dessas decisões. Nos últimos anos, ferramentas baseadas em dados e que auxiliam no processo de tomada de decisão estão sendo usadas para contratação (Miller, 2015), empréstimos bancários (Petrasic et al., 2017) e no sistema criminal (Barry-Jester et al., 2015). Essas decisões impactam a vidas das pessoas e por esse motivo há também um aumento da preocupação sobre o impacto que essas ferramentas podem causar na sociedade. Em particular, uma parcela da comunidade de aprendizado de máquina está focada em problemas relacionados à justiça, culpabilidade, transparência e ética destes sistemas (Beutel et al., 2019; Grgic-Hlaca et al., 2016; Saxena et al., 2018; Diakopoulos et al., 2017; Abebe et al., 2020). Com isso, a pesquisa por definições de justiça no contexto de aprendizado de máquina (AM) é uma das mais ativas na área.

Devido aos argumentos acima, ainda que algoritmos de AM atualmente auxiliam na tomada de decisão em contextos sócioeconômicos, os modelos treinados por estes algoritmos não são perfeitos. Isto é, erros podem ocorrer. Alguns exemplos são: COMPAS, uma ferramenta que pode condenar uma pessoa inocente por vieses raciais; aplicativo de fotos da Google identificando pessoas negras como gorilas; ferramenta de recrutamento da Amazon favorecendo homens para trabalhos técnicos; algoritmo de recorte de imagens do Twitter favorecendo homens brancos em vez de mulheres e pessoas negras. Em razão dessa possibilidade de erros, esses algoritmos devem ter responsabilidades junto a sociedade. Neste contexto, Diakopoulos (Diakopoulos et al., 2017) afirma que essas responsabilidades incluem *"uma obrigação de reportar, explicar ou justificar decisões algorítmicas assim como mitigar qualquer impacto negativo ou potencial prejuízo a sociedade"*.

Dada a relevância do tópico, foco dessa dissertação, múltiplas definições de justiça foram propostas na literatura da ciência da computação (Grgic-Hlaca et al., 2016; Saxena et al., 2018). Por exemplo, Narayanan (Narayanan, 2018) apresentou um tutorial detalhando 21 definições baseadas apenas em conceitos matemáticos e estatísticos sem abordar ou discutir a extensa literatura já existente na filosofia ou psicologia. Como consequência, não há um consenso entre pesquisadores sobre uma única definição de justiça (Gajane & Pechenizkiy, 2017). O problema é ainda mais complexo se considerarmos que as definições são incompatíveis entre si (Kleinberg et al., 2016).

Até o momento, a maior parte dos trabalhos no que diz respeito à justiça em aprendizado de máquina focaram no problema de aprendizado supervisionado usando dados que exploram casos criminais ou de empréstimo bancário (Grgic-Hlaca et al., 2016, 2018; Saxena et al., 2018). Em particular, a ferramenta usada apresentou problemas relacionados a racismo e preconceito de gênero (Chouldechova, 2017; Verma & Rubin, 2018).

Diante disso, essa dissertação é focada no levantamento e testes de hipótese relacionadas à aplicabilidade de conceitos de justiça em modelos de análise sobrevivência. Em análise de sobrevivência as modelagens estão interessadas em analisar o tempo até a ocorrência de um evento (e.g. morte de um paciente, estadia em uma cidade, falha de um componente mecânico). Para tal, exploramos não apenas diferentes modelos (CoxPH e CoxTime) como também bases em dois contextos socioeconômicos (saúde e criminal).

Até o momento não tivemos conhecimento de nenhum outro trabalho que tenha sugerido essa mesma proposta, o que demonstra uma enorme oportunidade de estudos e possíveis inovações para a área. A Figura 1.1 mostra um resumo de trabalhos anteriores em justiça nos diferentes domínios de aprendizado de máquina. É possível observar que não há nenhuma menção a análise de sobrevivência.

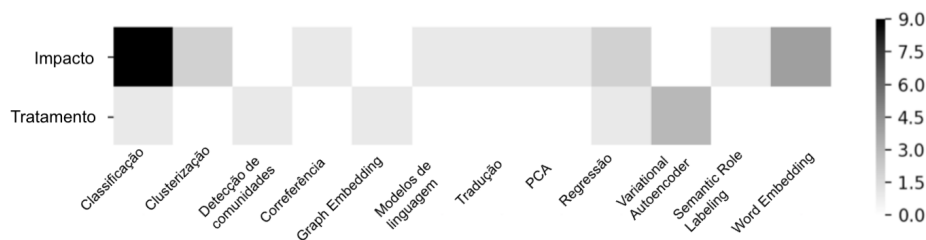


Figura 1.1. Heatmap com trabalhos anteriores em justiça agrupadas por domínio. Fonte: Adaptado de Mehrabi et al. (2019)

Para alcançar tal objetivo, essa dissertação é composta por cinco tópicos:

1. Levantamento da literatura em análise de sobrevivência (Capítulo 2)

2. Levantamento da literatura de justiça em aprendizado de máquina (Capítulo 3)
3. Adaptação das métricas de justiça para o contexto de modelos de sobrevivência (Capítulo 4)
4. Avaliação de diversos modelos de sobrevivência em bases com vieses de justiça (Capítulo 5)
5. Discussão das métricas, com foco em justiça, no modelo selecionado (Capítulo 6)

Como mencionado, um ponto central dessa dissertação é que abordamos três bases com características distintas. Uma é a base clássica Rossi, de reincidência criminal. A segunda é a COMPAS, uma base de dados que motivou essa discussão de justiça em AM. E por fim, uma base fora do contexto criminal, a MIMIC-III. Ao explorar três bases, podemos analisar aspectos diferentes de justiça. A base médica em particular é interessante pois apresenta novas complexidades.

Considere alguns exemplos. Doenças como câncer de pele terão maior incidência em pessoas com pele clara¹. Outro exemplo é câncer de mama ou próstata, que afeta mais um sexo do que o outro. Dadas essas características, ainda não é totalmente compreensível como pesquisadores conseguem quantificar justiça em serviços de saúde. Com isso, quando há uma relação causal entre um atributo sensível (por exemplo, sexo) e uma doença, é esperado que a falta de representatividade não seja um problema. No entanto, caso essa relação não seja presente, haverá um problema de representatividade, dado que é um fator importante para endereçar injustiças sociais e históricas (Saxena et al., 2018).

É importante entender essa diferença, uma vez que resultados discordantes para classes protegidas podem ser justas se motivadas por diferenças biológicas. Todavia, serão consideradas injustas caso os resultados tenham sido oriundos de decisões de tratamentos inadequados feitas no passado (Goodman et al., 2018). Lidar com justiça nesses casos pode ser complicado, uma vez que a acuracidade de sistemas de aprendizado de máquina irá depender de vários atributos entre eles variáveis sensíveis. Por outro lado, ao tomar decisões enviesadas (como por exemplo indicar a prioridade de atendimento em pessoas idosas), estará reforçando diferenças sociais e históricas que afetam diretamente a sociedade.

Um outro ponto relevante é que na adaptação de conceitos de justiça para métricas interpretáveis no mundo de modelos de análise de sobrevivência, no nosso trabalho nós propomos o uso de três abordagens. A primeira foca na disparidade entre curvas

¹<https://www.cancer.net/cancer-types/melanoma/statistics>

de sobrevivência observadas nos dados quando comparadas com previsões. Isto é, analisamos casos onde a curva de sobrevivência empírica extraída via Kaplan-Meier (Kaplan & Meier, 1958) apresenta uma diferença estatística entre dois grupos, porém a curva de predição do modelo não captura tal diferença. Chamamos tal métrica de divergência em paridade demográfica. Nossa segunda métrica consiste de um cálculo do c-index (Harrell et al., 1982) onde alteramos nos dados o grupo de interesse, por exemplo, um indivíduo que era do gênero masculino muda para feminino. Chamamos tal métrica de discriminação causal (Galhotra et al., 2017). Por fim, propomos uma métrica nova chamada de justiça de filas, onde comparamos cenários hipotéticos de duas pessoas sendo julgadas por um modelo de aprendizado de máquina ao mesmo tempo. Aqui buscamos discrepâncias entre o tempo observado do evento e a chance de sobrevivência. Um fator relevante das duas últimas métricas, discriminação causal e justiça de filas, é que diferente da primeira podemos analisar casos focados em cada indivíduo.

No próximo capítulo descrevemos os conceitos de modelos de sobrevivência para entender as novas métricas. Após isto, no Capítulo 3 descrevemos os conceitos de justiça. Finalmente, no Capítulo 4 detalhamos as métricas mencionadas acima. Como já dito, os capítulos seguintes focam nos nossos resultados (Capítulos 5 e 6). Por fim, Capítulo 7 fecha a dissertação.

Capítulo 2

Análise de sobrevivência

Neste capítulo iremos apresentar o conceito de análise de sobrevivência. Iniciamos com as definições e as notações que serão usadas por todo o capítulo. Em seguida são apresentados os principais métodos estatísticos descritos na literatura. Depois aprofundamos nos modelos de Cox e por fim destacamos as métricas utilizadas para análise de sobrevivência. Os conceitos envolvendo o modelo de Cox junto com as definições existentes de justiça, que serão apresentadas no Capítulo 3, irão servir como base para a proposta de novas definições focadas em modelos de análise de sobrevivência.

A análise de sobrevivência é um ramo da estatística responsável por modelar a ocorrência de eventos de interesse, tais como mortes, *churn*¹ de clientes ou a falha de um componente mecânico. O seu objetivo é fornecer respostas para perguntas do tipo: (i) Qual a probabilidade de sobrevivência de uma pessoa passado um tempo t ? (ii) Quais os efeitos que certas características de uma pessoa tem na sua probabilidade de sobrevivência? (iii) Qual a taxa de mortalidade da população estudada?

Em análise de sobrevivência, todos os pontos observados iniciam de um mesmo tempo t , podendo ser a internação em um hospital, início de um tratamento com certa medicação ou o diagnóstico de uma doença (Reddy et al., 2015). Considerando as várias aplicações deste tipo de análise, na prática o evento de interesse pode nunca ser observado, mas na área da saúde o mais usado é a morte biológica (Glazier, 2019). Sendo assim, é possível identificar o tempo exato do evento de interesse, porém há casos em que o evento não é observado durante o período de estudo e por isso ficaremos sem esse dado. Essa é uma característica importante em análise de sobrevivência e é conhecida como *censura*.

¹Termo usado para identificar clientes que cancelaram ou abandonaram um serviço prestado por uma empresa.

2.1 Definições

Na tabela 2.1 está definida a notação e terminologia que será usada em toda a dissertação para entender a definição formal de análise de sobrevivência. As definições serão apresentadas junto de um exemplo ilustrativo.

Antes de apresentar a notação segue algumas definições: vetores serão representados por letras minúsculas em negrito tais como \mathbf{x} e \mathbf{y} e valores numéricos serão representadas por letras minúsculas.

Tabela 2.1. Notações em análise de sobrevivência

Notações	Descrições
p	Número de atributos
n	Número de instâncias
\mathbf{x}	$\mathbb{R}^{n \times p}$ vetor de atributos
\mathbf{x}_i	$\mathbb{R}^{1 \times p}$ vetor de atributos da instância i
\mathbf{e}	$\mathbb{R}^{n \times 1}$ vetor dos eventos de interesse
\mathbf{c}	$\mathbb{R}^{n \times 1}$ vetor de censura
\mathbf{t}	$\mathbb{R}^{n \times 1}$ vetor de tempo da ocorrência do evento observado que é igual a $\min(\mathbf{e}, \mathbf{c})$
δ	$n \times i$ vetor binário que indica o estado do evento de interesse
β	$\mathbb{R}^{p \times 1}$ vetor de coeficientes
$f(t)$	Função densidade da distribuição de morte acumulada
$F(t)$	Função distributiva de morte acumulada
$S(t)$	Função de sobrevivência
$\lambda(t)$	Função de risco
$\Lambda(t)$	Função de risco acumulada

2.1.1 Censura

Durante o período de estudo algumas instâncias não terão seus eventos observados e esse conceito é definido como censura (Klein & Moeschberger, 2006). Isso pode acontecer devido a limitação da janela de tempo ou a vestígios perdidos, tais como pacientes que mudaram de hospital. Há três tipos de censura: (i) o primeiro e mais comum é a *censura à direita* - são consideradas as instâncias em que não ocorreu um evento de interesse durante o período observado. Por exemplo, um paciente que opta por deixar o estudo antes do final e a morte não é observada. (ii) *censura à esquerda*

- Não se conhece o momento do evento mas sabemos que ele ocorreu antes do tempo observado. (iii) *Censura-intervalar* - O evento de interesse ocorre em um intervalo de tempo mas o mesmo não é especificado. (Glazier, 2019). Nessa dissertação iremos trabalhar com o tipo (i) censura à direita, visto que as bases escolhidas possuem esse tipo de censura.

Em análise de sobrevivência, além do **evento de interesse**, que é um vetor binário composto por 1's e 0's, onde 1 é um evento observado e 0 um evento não observado, também temos o vetor do *tempo do evento de interesse* (\mathbf{t}). O tempo é conhecido apenas para as instâncias em que o evento ocorreu durante o período observado, para todas as outras instâncias o tempo do evento de interesse é maior que o período de estudo e por isso só temos o tempo de censura (\mathbf{c}). Com isso, para qualquer instância i , é possível observar o tempo de sobrevivência (\mathbf{t}_i) ou o tempo de censura (\mathbf{c}_i).

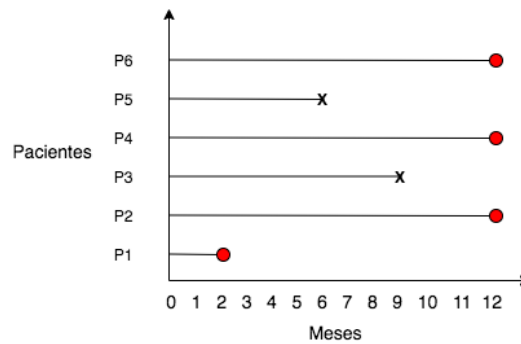


Figura 2.1. Ilustração demonstrando um problema de análise de sobrevivência. Fonte: adaptado de Wang et al. (2019)

Na Figura 2.1 é dado um exemplo ilustrativo de um problema de análise de sobrevivência. Nesse exemplo, são observados seis pacientes por um período de doze meses. Pela Figura 2.1 é possível observar que o paciente P1 deixou o estudo (seja por alta hospitalar ou transferência para outro hospital). Para os pacientes P2, P4 e P6 não houve evento de interesse durante os 12 meses, portanto eles são considerados censurados. E para os pacientes P3 e P5 houve evento de interesse (eg. morte) e o tempo observado é o tempo do evento de interesse.

Declaração do problema: Para uma dada instância i , representada pela tupla $(\mathbf{x}_i, \mathbf{t}_i, \delta_i)$, onde \mathbf{x}_i é o vetor de atributos da instância i ; δ_i é indicador do estado do evento, isto é, $\delta_i = 1$ para instâncias com eventos não censurados e $\delta_i = 0$ para instâncias com eventos censurados; e \mathbf{t}_i representa o tempo observado: \mathbf{e}_i se a instância não for censurada e \mathbf{c}_i caso seja censurada, isto é,

$$\mathbf{t}_i = \begin{cases} \mathbf{e}_i & \text{se } \delta_i = 1 \\ \mathbf{c}_i & \text{se } \delta_i = 0 \end{cases} \quad (2.1)$$

Com isso, o principal objetivo da análise de sobrevivência é estimar o tempo da ocorrência do evento de interesse \mathbf{t}_j para uma nova instância j a partir do vetor de atributos \mathbf{x}_j (Wang et al., 2019).

2.1.2 Função de sobrevivência

A função de sobrevivência ($S(t)$) é definida como

$$S(t) = P(T > t) \quad (2.2)$$

onde P é a probabilidade de sobrevivência, t é um tempo arbitrário e T é uma variável aleatória representando o tempo da ocorrência do evento. Logo, para um dado tempo t , a função retorna a probabilidade de sobrevivência passado esse tempo t . Assuma-se que $S(0) = 1$, portanto todos os pacientes estão vivos e não ocorreu um evento de interesse. Um fato importante é que $S(t)$ não aumenta com o tempo t , ou seja, $S(a) \leq S(t)$, $\forall a \geq t$ (Glazier, 2019).

A partir da função de sobrevivência pode-se calcular o complemento dela, também conhecida como **função distributiva de morte acumulada** e é definida como

$$F(t) = 1 - S(t) \quad (2.3)$$

Essa função retorna a probabilidade que o evento já tenha ocorrido dado um tempo t . Considerando que a função é diferenciável, obtém-se a **função densidade da distribuição de morte acumulada**, que é definida como

$$\begin{aligned} f(t) &= F'(t) && \text{para casos contínuos} \\ f(t) &= \frac{F(t + \Delta t) - F(t)}{\Delta t} && \text{para casos discretos} \end{aligned}$$

e nada mais é do que a taxa de eventos de interesse por unidade de tempo. Com isso temos que as funções de sobrevivência, distributiva e densidade estão relacionadas e também podem ser definidas como:

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du = 1 - F(t) \quad (2.4)$$

e a Figura 2.2 apresenta a relação entre essas funções.

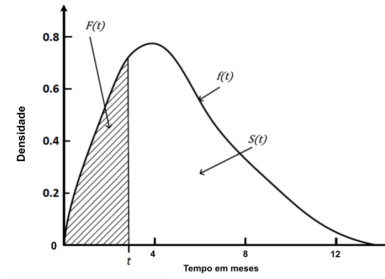


Figura 2.2. Ilustração apresentando relações entre as funções $S(t)$, $F(t)$ e $f(t)$. Fonte: adaptado de Wang et al. (2019)

2.1.3 Função de risco

A função de risco ($\lambda(t)$) pode ser definida como a taxa instantânea de ocorrência do evento de interesse no tempo t para os indivíduos que sobreviveram até o tempo t . Matematicamente, a função é definida como

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t \cdot S(t)} = \frac{f(t)}{S(t)} \quad (2.5)$$

Essa função é não negativa, assim como a $S(t)$, e pode ter uma variedade de formas desde que as seguintes propriedades sejam satisfeitas:

- $\lambda(t) \geq 0, \forall x \geq 0$
- $\int_0^{\infty} \lambda(t) dt = \infty$

Um exemplo de função de risco é a curva da banheira, mostrada na Figura 2.3. Nessa imagem, para valores menores de t a função está decrescendo até atingir um mínimo; ela permanece estável por um tempo e após começa a crescer novamente. Esse é o caso de peças em sistemas mecânicos e também de mortalidade em seres humanos: no início o risco de falha ou morte é alto, depois se estabiliza por anos, e volta a crescer por desgaste ou devido a velhice.

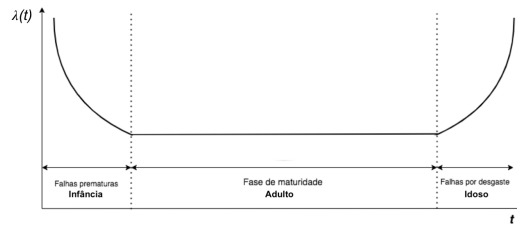


Figura 2.3. Curva da banheira

Considerando que a $f(t)$ também pode ser definida como $-\frac{d}{dt}S(t)$, então a $\lambda(t)$ fica:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}S(t) \cdot \frac{1}{S(t)} = -\frac{d}{dt}[\ln S(t)] \quad (2.6)$$

e portanto a equação 2.2 pode ser reescrita como:

$$S(t) = \exp(-\Lambda(t)) \quad (2.7)$$

onde $\Lambda(t) = \int_0^t \lambda(u)du$ é a função de risco acumulada (Lee & Wang, 2003).

Todas essas equações são diferentes maneiras de representar a mesma distribuição de sobrevivência no tempo t . A partir de uma deriva-se todas as demais. Isso é importante pois torna a análise de sobrevivência mais interpretável e acessível.

2.2 Métodos estatísticos tradicionais

Devido a grande variedade de populações e atributos que podem afetar o tempo de sobrevivência de cada indivíduo, é necessário ter diferentes métodos para modelar cada problema. Nessa seção serão apresentados os três tipos de métodos para estimar as funções de sobrevivência e/ou de risco: (i) não-paramétricas, (ii) semi-paramétricas e (iii) paramétricas.

2.2.1 Métodos não-paramétricos

São métodos que não fazem suposições sobre a distribuição $S(t)$ e portanto são mais eficientes quando usados em problemas que não se tem uma distribuição conhecida. Nesse caso, os métodos geram uma estimativa empírica da curva de sobrevivência a partir dos dados. Um dos métodos mais conhecidos é o de Kaplan-Meier.

2.2.1.1 Kaplan-Meier

A curva Kaplan-Meier (KM), também conhecida como estimador limite-produto, foi desenvolvida em 1958 por Kaplan e Meier (Kaplan & Meier, 1958) e utiliza o tamanho real do tempo observado para estimar a função de sobrevivência.

O estimador Kaplan-Meier para a função de sobrevivência, que é o produto das probabilidades de sobreviver além do tempo t , é definida como:

$$S(t) = \prod_{i:t_i \leq t} \left(1 - \frac{e_i}{n_i}\right) \quad (2.8)$$

onde o t_i é o tempo no qual pelo menos um evento ocorreu, e_i é o número de eventos que ocorreu no tempo t_i e n_i é o número de indivíduos que sobreviveram até o tempo t_i . Importante salientar que para obter n_i os indivíduos censurados são considerados no cálculo, visto que eles tem um efeito na curva de sobrevivência mesmo que seus eventos não tenham sido observados. Logo para calcular n_i é necessário considerar n_{i-1} , e_{i-1} e c_{i-1} no tempo t_{i-1} , onde c_{i-1} são os indivíduos censurados.

A Figura 2.4 ilustra um exemplo de uma curva KM usando os dados de reincidência da base de dados Rossi (Rossi et al., 1980).

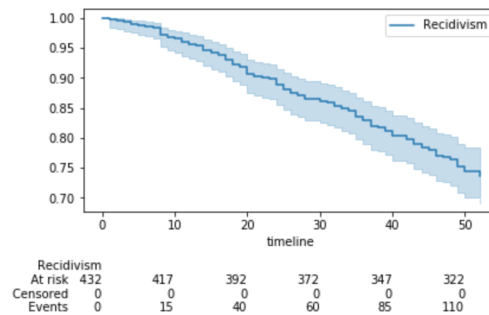


Figura 2.4. Exemplo de uma curva Kaplan-Meier

Um uso muito comum desta função é em análises que querem comparar as curvas de sobrevivência de duas sub-populações. Por exemplo, em um teste clínico, pesquisadores querem comparar o efeito de duas drogas experimentais A e B no tempo de sobrevivência das populações (Glazier, 2019). No entanto, caso o número de indivíduos ou a população seja muito grande é recomendado usar um outro método chamado Tabela de Vida (Cutler & Ederer, 1958) mas que não será discutido nessa dissertação.

2.2.2 Métodos paramétricos

Os métodos paramétricos são utilizados quando os dados possuem uma distribuição conhecida. Entre as mais comuns estão: exponencial, Weibull, logística e log-logística (Wang et al., 2019).

Tabela 2.2. Funções densidade, sobrevivência e de risco para distribuições que usam métodos paramétricos. Fonte: Wang et al. (2019)

Distribuição	Densidade $f(t)$	Sobrevivência $S(t)$	Risco $\lambda(t)$
Exponencial	$b \exp(-bt)$	$\exp(-bt)$	b
Weibull	$(\frac{k}{b})(\frac{t}{b})^{k-1} \exp(-\frac{t}{b})^k$	$\exp(-(\frac{t}{b})^k)$	$(\frac{k}{b})(\frac{t}{b})^{k-1}$
Logística	$\frac{\exp(-\frac{(t-\mu)}{\sigma})}{\sigma(1+\exp(-\frac{(t-\mu)}{\sigma}))^2}$	$\frac{\exp(-\frac{(t-\mu)}{\sigma})}{1+\exp(-\frac{(t-\mu)}{\sigma})}$	$\frac{1}{\sigma(1+\exp(-\frac{(t-\mu)}{\sigma}))}$
Log-Logística	$\frac{bkt^{k-1}}{(1+bt^k)^2}$	$\frac{1}{1+bt^k}$	$\frac{bkt^{k-1}}{1+bt^k}$

2.2.2.1 Distribuição exponencial

Este modelo é o mais simples entre os demais e a sua função de risco possui apenas um parâmetro b . Um alto valor de b indica um alto risco de ocorrência do evento de interesse e conseqüentemente uma curva de sobrevivência menor (Wang et al., 2019).

2.2.2.2 Distribuição Weibull

Este modelo é o mais usado para o problema de sobrevivência entre os métodos paramétricos. Ele possui dois parâmetros $b > 0$ e $k > 0$ e a função de risco terá a sua forma definida pelo valor de k . Caso $k = 1$ a função de risco será constante e o modelo será igual ao modelo exponencial. Caso $k < 1$ a função de risco irá diminuir com o tempo t e a proporção será determinada pelo parâmetro b .

2.2.2.3 Distribuições logística e log-logística

No modelo logístico há dois parâmetros: o parâmetro de localização, $\mu \in \mathbb{R}$, e o parâmetro de escala, $\sigma > 0$. Dependendo do valor de μ a função desloca-se ao longo do eixo das abcissas.

Para o modelo log-logístico, o parâmetro $k > 0$ irá determinar o formato da curva. Caso o $k > 1$ a função de risco irá crescer até o máximo para depois decrescer no tempo t . Caso $k \leq 1$ a função só irá decrescer no tempo t .

A tabela 2.2 sumariza através das funções densidade, sobrevivência e de risco, as distribuições exponencial, Weibull, logística e log-logística discutidas nessa seção.

2.2.3 Métodos semi-paramétricos

Como o próprio nome diz, os métodos semi-paramétricos são compostos de funções paramétricas e não-paramétricas. Esses modelos podem obter um estimador mais consistente quando comparado com métodos paramétricos e um estimador mais preciso quando comparado com métodos não-paramétricos (Powell, 1994). Para esses métodos não é necessário conhecer a distribuição dos eventos de interesse, porém assume-se que os atributos tem uma influência exponencial no resultado. Nessa categoria o mais usado é o modelo de Cox (Cox, 1972).

Dada que nossa maior ênfase é no modelo de Cox, iremos discutir o mesmo em detalhes na próxima sub-seção.

2.2.4 Cox de riscos proporcionais

Esse é o modelo clássico e mais básico cuja a função de risco possui a seguinte forma:

$$\lambda(t, \mathbf{x}_i) = \lambda_0(t) \exp(\mathbf{x}_i \boldsymbol{\beta}) \quad (2.9)$$

onde $i = 1, 2, 3, \dots, n$ são as instâncias, \mathbf{x}_i é o vetor de atributos da instância i , $\boldsymbol{\beta}$ é o vetor de coeficientes, que quantifica o efeito de cada atributo no risco de ocorrer um evento de interesse. $\lambda_0(t)$ é a função de risco base, que é uma função arbitrária e não-negativa.

Se compararmos duas instâncias \mathbf{x}_1 e \mathbf{x}_2 , teremos uma taxa de risco dada por:

$$\frac{\lambda(t, \mathbf{x}_1)}{\lambda(t, \mathbf{x}_2)} = \frac{\lambda_0(t) \exp(\mathbf{x}_1 \boldsymbol{\beta})}{\lambda_0(t) \exp(\mathbf{x}_2 \boldsymbol{\beta})} = \exp((\mathbf{x}_1 - \mathbf{x}_2) \boldsymbol{\beta}). \quad (2.10)$$

A equação acima mostra que essa taxa é independente da função de risco base. Com isso, o modelo de cox é considerado semi-paramétrico pois a função $\lambda_0(t)$ não é especificada e é chamado de risco proporcional pois a taxa de risco é constante para todas as instâncias, como provado na equação 2.10. Além disso, todas as instâncias usam a mesma função de risco base. Se integrarmos a fórmula 2.9 dos dois lados em relação a t , teremos:

$$\Lambda(t, \mathbf{x}) = \Lambda_0(t) \exp(\mathbf{x} \boldsymbol{\beta}) \quad (2.11)$$

que é a função de risco acumulada. Se combinarmos essa equação com a 2.7 teremos a equação de sobrevivência:

$$S(t) = \exp(-\Lambda_0(t) \exp(\mathbf{x}\boldsymbol{\beta})) = S_0(t)^{\exp(\mathbf{x}\boldsymbol{\beta})}. \quad (2.12)$$

aqui, $\Lambda_0(t)$ é a função base de risco acumulado e $S_0(t) = \exp(-\Lambda_0(t))$ representa a função *baseline* de sobrevivência.

Como a função de risco base $\lambda_0(t)$ no modelo de Cox não é especificada não é possível ajustar o modelo usando a função de verossimilhança tradicional. No entanto, o que realmente se tem interesse é no vetor de coeficientes ($\boldsymbol{\beta}$). Para isso, Cox propôs a função de verossimilhança parcial (Cox, 1975) que depende apenas do vetor de parâmetros $\boldsymbol{\beta}$, tal que:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \lambda(t_i)^{\delta_i} S(t_i) \quad (2.13)$$

onde δ_i é um indicador de evento, assumindo o valor um se ocorreu o evento para o indivíduo i e zero caso contrário. O objetivo é estimar $\hat{\boldsymbol{\beta}}$ maximizando a função de verossimilhança parcial.

Uma variação do modelo se chama Cox-Time. Tal variação foi proposta por Kvamme et al. (2019) e procura solucionar as restrições do modelo clássico. Para isso foi definida uma nova função de risco dependente do tempo e com isso deixa de ser proporcional:

$$\lambda(t, \mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}\boldsymbol{\beta}t) \quad (2.14)$$

2.3 Métricas de Avaliação

Devido à natureza temporal do modelo, além da presença de dados censurados, para avaliar modelos de sobrevivência não podemos possível utilizar as métricas tradicionais de modelos lineares: como a raiz do erro quadrático médio ou o coeficiente de determinação (R^2). Nesta seção, discutimos as métricas de qualidade de ajuste e previsão particulares para os modelos utilizados na dissertação.

2.3.1 C-Index

O índice de concordância, também conhecido como c-index (Harrell et al., 1982), estima a probabilidade de pares aleatórios terem as curvas de sobrevivência preditas com as mesmas ordens que as suas curvas verdadeiras (Glazier, 2019). Isto é, para cada dois pacientes escolhidos aleatoriamente, é dito que eles são concordantes se o paciente

com maior risco predito tem um tempo de sobrevivência menor que o outro paciente. Além disso, as curvas de sobrevivência só podem ser comparadas em dois cenários: (i) ambos eventos não são censurados e (ii) o tempo do evento observado deve ser menor que a do evento censurado (Wang et al., 2019).

Algumas propriedades importantes são:

- O valor do c-index varia entre 0 e 1, sendo 1 a concordância perfeita e 0 a anti-concordância.
- Um valor de 0.5 é considerado um resultado aleatório (Harrell et al., 1982).
- Em geral, modelos bem treinados possuem um resultado que varia entre 0.6 e 0.75 devido a presença de ruídos nos dados que dificultam os modelos de serem perfeitamente concordantes (Glazier, 2019).

Um ponto relevante é que o c-index por ser uma métrica que está mais preocupada com a ordem dos pares, não captura a corretude das curvas preditas. Por isso é indicado fazer uma verificação além do resultado do c-index. Porém, se as curvas estiverem corretas então elas terão consequentemente um bom c-index (Glazier, 2019).

Outro detalhe é que o c-index serve muito bem para os modelos de Cox de riscos proporcionais, pois a ordem não será alterada no tempo. No entanto, isso não se mantém para o caso do Cox Time, onde não temos mais um modelo de risco proporcional. Neste caso é necessário usar uma métrica desenvolvida por Antolini et al. (2005) que é dependente do tempo com uma pequena variação proposta por Kvamme et al. (2019):

$$C_{td} = \frac{\sum_{i,j} 1_{\hat{S}(t_i|\mathbf{x}_i) < \hat{S}(t_j|\mathbf{x}_j)} \cdot 1_{t_i < t_j} \cdot \delta_i}{\sum_{i,j} 1_{t_i < t_j} \cdot \delta_i} \quad (2.15)$$

No mais, a ideia é a mesma, a probabilidade de sobrevivência de um indivíduo i será menor que a de um indivíduo j dado que tempo de ocorrência do evento de i é anterior ao de j e o evento ocorreu para o indivíduo i . O c-index tem uma relação próxima com a métrica de acurácia para modelos de classificação (Ishwaran et al., 2008) e a curva ROC (Heagerty & Zheng, 2005).

2.3.2 IPCW Brier Score

A *inverse probability of censoring weighted Brier Score* é uma métrica usada em modelos que possuem saídas probabilísticas dentro da faixa $[0,1]$ cujo o somatório final é um e que serve para avaliar as estimativas do modelo. A definição do Brier Score dado um tempo t é:

$$BS(t) = \frac{1}{n} \sum_{i=1} [1 - \hat{S}(t_i)]^2 \quad (2.16)$$

onde n é o número de indivíduos e $\hat{S}(t_i)$ é a probabilidade de sobrevivência para o indivíduo i . Essa métrica foi ampliada por Graf et al. (1999) para comportar dados censurados e passou a ser definida como:

$$BS(t) = \frac{1}{n} \sum_{i=1} \frac{(-\hat{S}(t_i))^2}{G(C)} + \frac{(1 - \hat{S}(t_i))^2}{G(t)}. \quad (2.17)$$

Aqui, a ideia é atribuir um peso a instância i . Esse peso é obtido a partir do estimador de Kaplan-Meier para a distribuição censurada, tal que $G(t) = P(C > t)$, onde C é uma variável aleatória representando o tempo da censura:

A partir dessa distribuição com pesos, as instâncias censuradas antes do tempo t terão valor zero, porém continuam a contribuir para o cálculo pois foram usadas para calcular a função G . Já os pesos das instâncias não censuradas no tempo t possuem valor maior que zero e portanto contribuem para o cálculo do Brier Score.

Neste capítulo apresentamos os conceitos de análise de sobrevivência e o modelo de Cox. No Capítulo 3 vamos apresentar os conceitos de aprendizado de máquina e as principais definições de justiça. Esses dois capítulos serão usados como base para as propostas apresentadas no Capítulo 4.

Capítulo 3

Justiça em aprendizado de máquina

Neste capítulo serão apresentados as definições de aprendizado de máquina (AM) e justiça em aprendizado de máquina. Em particular, na Seção 3.1 começamos com uma visão geral sobre AM, em especial problemas de classificação, e justiça. Em seguida aprofundamos no conceito de justiça e por fim, na Seção 3.3 apresentamos as principais definições de justiça em AM.

3.1 Visão geral sobre aprendizado de máquina e justiça

Aprendizado de máquina (AM) pode ser definido como um conjunto de métodos usados para automaticamente descobrir padrões nos conjuntos de dados, e então usar esses padrões para fazer previsões em dados futuros. Uma maneira de resolver problemas envolvendo incerteza é usando a teoria da probabilidade, que em AM pode ter várias formas: qual a melhor previsão sobre o futuro de acordo com dados passados? Qual o melhor modelo que representa esses dados? (Murphy, 2012).

Embora existam várias abordagens para AM, nessa dissertação iremos focar no tipo de problema mais explorado em justiça, que é o aprendizado supervisionado (AS), logo vamos aprofundar esse conceito nessa seção. No caso de AS, o problema envolve aprender a prever qual a saída correta a partir de novos dados e tomar decisões mediante incerteza.

3.1.1 Visão geral sobre aprendizado supervisionado

Nessa classe de problemas, o objetivo é aprender como mapear as entradas \mathbf{x}_i para as saídas \mathbf{y}_i . Formalmente temos que, dado um conjunto de treinamento \mathcal{D} com

n pares de entrada-saída:

$$\mathcal{D} = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n \quad (3.1)$$

onde \mathbf{x}_i é um vetor de atributos e cada atributo deste vetor é indexado por j e portanto $\mathbf{x}_{ij} \in \mathbb{R}$ e $\mathbf{y}_i \in \mathbb{R}$. Em que \mathbf{y}_i foi gerado por uma função desconhecida $\mathbf{y}_i = f(\mathbf{x}_i)$. Sendo assim, o objetivo é descobrir a função h que mais se aproxima da verdadeira função f (Russell & Norvig, 2016). Logo, no aprendizado supervisionado, o propósito é aprender a função $\hat{\mathbf{y}}_i = h_{\Theta}(\mathbf{x}_i)$ que com maior exatidão generaliza o conjunto de dados \mathcal{D} . Aqui, $\hat{\mathbf{y}}_i$ é uma predição. Para verificar se o modelo está generalizando bem é necessário utilizar um outro conjunto de dados, chamado de conjunto teste, que terá novos pares de entrada-saída. Importante salientar que nesta dissertação iremos considerar apenas saídas binárias, ou seja, $\hat{\mathbf{y}}_i \in \{0, 1\}$.

3.1.2 Exemplo de uma aplicação real

Correctional Offender Management Profiling for Alternative Sanction (COMPAS) é um algoritmo que foi desenvolvido pela empresa Northpointe (hoje chamada de Equivant¹) para prever a chance de uma pessoa que foi presa em reincidir, ou seja, repetir um crime ou delito. As entradas \mathbf{x}_i são derivadas de 137 perguntas que são respondidas pelos presos ou retiradas de registros criminais e a saída \mathbf{y}_i é um pontuação que indica se a pessoa tem alto ou baixo risco de reincidir.

O COMPAS é usado pela justiça em vários estados dos Estados Unidos (EUA) há vários anos (Kirkpatrick, 2017), porém possui um gravíssimo problema de enviesamento contra pessoas negras. Para ilustrar esse problema, veja as tabelas presentes na Figura 3.1.

	réus negros		réus brancos	
	risco alto	risco baixo	risco alto	risco baixo
reincidiu	1369	532	505	461
não reincidiu	805	990	349	1139

Figura 3.1. Tabelas com dados do Compas. Fonte: Adaptado de (Gummadi, 2019)

Na tabela da esquerda, temos a quantidade de réus negros que foram classificados pelo modelo como risco alto e reincidiram, risco alto e não reincidiram, risco baixo e

¹<https://www.equivant.com/>

reincidiram e risco baixo e não reincidiram. Na tabela da direita, temos a mesma lógica, só que para réus brancos. Podemos interpretar cada tabela como uma matriz de confusão e calcular algumas métricas. O problema aparece quando calculamos as taxas de falsos positivos (TFP) e falsos negativos (TFN):

$$\begin{aligned} TFP_{negros} &= \frac{FP}{FP+VN} = \frac{805}{805+990} = 0,45 & TFN_{negros} &= \frac{FN}{FN+VP} = \frac{532}{532+1369} = 0,29 \\ TFP_{brancos} &= \frac{FP}{FP+VN} = \frac{349}{349+1139} = 0,23 & TFN_{brancos} &= \frac{FN}{FN+VP} = \frac{461}{461+505} = 0,48 \end{aligned}$$

A taxa de falsos positivos é muito maior para as pessoas negras do que para as brancas enquanto a de falsos negativos é menor para pessoas negras do que para brancas. Isso diz que pessoas negras que não voltaram a cometer crimes foram classificadas como alto risco de reincidir enquanto pessoas brancas que foram classificadas com baixo risco de reincidir voltaram a cometer crimes. Do ponto de vista prático, mais negros continuaram presos enquanto mais brancos foram soltos.

A Pro Publica fez uma reportagem² em 2016 mostrando como o COMPAS é usado e os impactos que essa aplicação está tendo na sociedade dos EUA.

Com base no exemplo acima e através da matriz de confusão, podemos ver que é possível definir várias métricas de justiça (por exemplo, diferentes normalizações). Antes de entrar em tais métricas, vamos aprofundar no conceito de justiça na próxima seção.

3.2 Justiça

Um foco recente da comunidade de aprendizado de máquina é sobre justiça nesses modelos. Como discutido no Capítulo 1, no contexto de tomada de decisão, justiça pode ser considerada a ausência de preconceitos ou favoritismos em relação a um indivíduo ou grupo de indivíduos baseado em suas características intrínsecas ou adquiridas (Mehrabi et al., 2019).

No entanto, visto que muitos sistemas agora são *data-driven*, essas decisões podem afetar negativamente a vida de seres humanos, reforçando desigualdades sociais e históricas. Por isso, pesquisar sobre justiça aplicada a modelos de aprendizado de máquina é extremamente importante e uma das áreas mais ativas atualmente.

Desde 2011, o número de definições sobre justiça cresce. Até hoje, aproximadamente 20 definições existem e cada uma possui detalhes e diferenças que tornam complexa a coexistência delas na mesma situação. Por esse motivo, nós investiga-

²<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

mos várias definições de justiça (Gajane & Pechenizkiy, 2017; Grgic-Hlaca et al., 2016; Narayanan, 2018; Verma & Rubin, 2018) e que serão aprofundadas nessa dissertação.

Em particular, vamos explorar os conceitos de *discriminação por impacto desigual* e *tratamento desigual*. Tais são conceitos legais elaborados nos Estados Unidos que guiam o entendimento de julgamentos justos e injustos.

3.2.1 Discriminação por tratamento desigual e discriminação por impacto desigual

Na Seção 3.1 mostramos que os sistemas de aprendizado de máquina usam métricas que focam em minimizar os erros a partir dos dados históricos. No entanto, é possível que classificadores que estejam com os resultados otimizados tomem decisões para grupos sociais diferentes e que essas métricas possuam diferentes resultados para esses grupos.

Em vários países há leis anti-discriminatórias que proíbem o tratamento injusto de pessoas baseado em características específicas, chamadas de atributos sensíveis (p.ex. raça, gênero). Essas leis geralmente utilizam os conceitos de *discriminação por tratamento desigual* e *discriminação por impacto desigual* (Barocas & Selbst, 2016). O primeiro é usado quando as decisões mudam em razão dos atributos sensíveis de uma pessoa e o segundo é usado quando o resultado beneficia ou machuca desproporcionalmente indivíduos de um certo grupo (Zafar et al., 2017).

Um exemplo de *discriminação por tratamento desigual* ocorre quando uma empresa deixa de contratar uma pessoa qualificada para a posição em aberto devido a características raciais, de gênero ou por deficiência. **Phillips v. Martin Marietta Corp.**³ foi um caso em 1971 nos Estados Unidos, onde Ida Phillips não foi contratada para uma vaga na empresa Martin Marietta Corporation por ser mãe de uma criança pequena, enquanto essa regra não se aplicava para homens.

Um exemplo de *discriminação por impacto desigual* ocorre quando uma empresa avaliando candidatos para uma vaga faz as mesmas exigências para todos, porém alguma das condições acaba prejudicando um dos grupos. **Griggs et al. v. Duke Power Co.**⁴ foi um dos primeiros casos conhecidos de *discriminação por impacto desigual*, onde a Suprema Corte dos Estados Unidos julgou que a empresa estava exigindo requisitos dos candidatos que não eram pertinentes ao trabalho que eles iriam desempenhar.

³https://en.wikipedia.org/wiki/Phillips_v._Martin_Marietta_Corp.

⁴https://en.wikipedia.org/wiki/Griggs_v._Duke_Power_Co.

Na década de 50, a empresa Duke Power's Dan River Steam Station tinha uma política de não permitir negros em um dos seus departamentos. A partir de 1955 eles passaram a exigir diploma de ensino médio para todos os departamentos, exceto o que tinha somente pessoas brancas, e em 1965 passou a exigir testes de aptidão mecânica e QI para quem quisesse mudar para departamentos que pagassem mais.

Porém, de acordo com o Censo de 1960 nos EUA, 34% dos homens brancos na Carolina do Norte tinham diploma de ensino médio, em comparação com apenas 18% dos negros. A disparidade era ainda maior em relação as notas dos testes, onde 58% dos brancos passavam em comparação com somente 6% dos negros. Logo, foi julgado que a empresa estava involuntariamente discriminando contra negros.

A partir desses dois conceitos foram criadas definições em aprendizado de máquina que pudessem representar o significado legal. Importante salientar que a lista da tabela 3.1 não é exaustiva e portanto não contém todas as definições existentes na literatura.

Tabela 3.1. Visão geral das definições de justiça

Definição	Tratamento	Impacto
Paridade demográfica / estatística		✓
Paridade demográfica condicionada		✓
Discriminação causal	✓	
Oportunidade equivalente		✓
Probabilidade equiparada		✓
Tratamento por equidade		✓
Teste de justiça / calibração		✓
Justiça por conhecimento	✓	
Justiça por desconhecimento	✓	
Justiça por contrafactual	✓	

3.3 Principais definições de justiça

Nessa seção iremos apresentar cada uma das definições presentes na tabela 3.1. Para cada uma das definições que serão apresentadas assuma que o indivíduo ou grupo possua uma variável sensível z e que ela seja binária, ou seja, $z \in \{0, 1\}$ e que ela possua um vetor de atributos \mathbf{x}_j . A tabela 3.2 apresenta a notação que será usada nas definições de justiça presentes nessa seção.

Tabela 3.2. Notação para classificação binária

Notação	Descrição
$y \in \{0, 1\}$	Saída com valor verdadeiro
$\hat{y} \in \{0, 1\}$	Saída com valor predito
$s = p(\hat{y} = 1)$	Pontuação predita. Probabilidade de $\hat{y} = 1$.
$z \in \{0, 1\}$	Variável sensível
\mathbf{x}_j	Vetor de atributos, onde $j=1,\dots,n$

- **Paridade demográfica / estatística**

Se diz que um classificador binário satisfaz essa definição caso:

$$p(\hat{y} = 1|z = 0) = p(\hat{y} = 1|z = 1), \quad (3.2)$$

que é a probabilidade do classificador atribuir a classe positiva, $\hat{y} = 1$, para ambos os valores da variável sensível z (Dwork et al., 2012). No exemplo do COMPAS seria a mesma probabilidade de atribuir um risco alto para negros e brancos.

- **Paridade demográfica condicionada**

Essa definição expande a anterior permitindo que um conjunto de atributos \mathbf{w}_j influencie a resposta. Se diz que um classificador satisfaz essa definição caso:

$$p(\hat{y} = 1|\mathbf{w}_j, z = 0) = p(\hat{y} = 1|\mathbf{w}_j, z = 1), \quad (3.3)$$

que é a probabilidade do classificador atribuir a classe positiva, $\hat{y} = 1$, para ambos os valores da variável sensível z , controlando por um conjunto de atributos \mathbf{w}_j permitidos (Corbett-Davies et al., 2017), onde $\mathbf{w}_j \subset \mathbf{x}_j$. Um exemplo desses atributos, no caso de solicitação de crédito pode ser o emprego atual da pessoa e o histórico de crédito (Verma & Rubin, 2018). No exemplo do COMPAS, a probabilidade de atribuir um risco alto para negros e brancos, dado a idade e histórico de violência como fatores que influenciam a resposta, seria a mesma.

- **Discriminação causal**

Um classificador binário satisfaz essa definição caso:

$$p(\hat{y} = 1|z = 0, \mathbf{x}_{aj}) = p(\hat{y} = 1|z = 1, \mathbf{x}_{bj}), \mathbf{x}_{aj} \sim \mathbf{x}_{bj} \quad (3.4)$$

ou seja, se produzir o mesmo resultado para quaisquer dois indivíduos a e b que possuam um conjunto de atributos \mathbf{x}_j muito parecidos entre si mas que diferem na variável sensível z (Galhotra et al., 2017). No exemplo do COMPAS, a probabilidade de atribuir um risco alto para uma pessoa negra e outra branca que possuem atributos similares seria a mesma.

- **Oportunidade equivalente**

Um classificador binário satisfaz essa definição quando

$$p(\hat{y} = 1|z = 0, y = 1) = p(\hat{y} = 1|z = 1, y = 1) \quad (3.5)$$

que é a probabilidade de uma pessoa que possui uma classe positiva ser corretamente atribuída como positiva para ambos os valores da variável sensível z (Hardt et al., 2016). No exemplo do COMPAS, a probabilidade de atribuir um risco alto para negros e brancos que de fato o possuem, seria a mesma.

- **Probabilidade equiparada**

Um classificador binário satisfaz essa definição caso:

$$p(\hat{y} = 1|z = 0, y = i) = p(\hat{y} = 1|z = 1, y = i), i \in \{0, 1\} \quad (3.6)$$

Isso significa que a probabilidade de uma pessoa da classe positiva ser corretamente predita como positiva e a probabilidade de uma pessoa da classe negativa ser incorretamente predita como positiva é a mesma para ambos os valores da variável sensível z (Hardt et al., 2016). No exemplo do COMPAS, a probabilidade de corretamente atribuir um risco alto para quem de fato é e a probabilidade de incorretamente atribuir um risco alto para quem não é, é a mesma para pessoas negras e brancas.

- **Tratamento por equidade**

Um classificador binário satisfaz essa definição caso:

$$\frac{FN}{FP}_{z=0} = \frac{FN}{FP}_{z=1} \quad (3.7)$$

tal que a razão de falsos negativos sobre falsos positivos seja a mesma para ambos os valores da variável sensível z (Berk et al., 2018). No exemplo do COMPAS, a taxa de falsos negativos dividido por falsos positivos é a mesma para pessoas negras e brancas.

- **Teste de justiça / calibração**

Um classificador binário satisfaz essa definição caso:

$$p(y = 1|s, z = 0) = p(y = 1|s, z = 1) \quad (3.8)$$

isso significa que para qualquer valor de s predito, a probabilidade de realmente pertencer a classe positiva é a mesma para ambos os valores da variável sensível z (Chouldechova, 2017). No caso do COMPAS, a probabilidade de reincidir, para todos os valores de s , é a mesma para pessoas negras e brancas.

- **Justiça por conhecimento**

O princípio aqui é que indivíduos com atributos similares devem ser classificados de forma similar. Essa similaridade é definida por uma métrica de distância; para a definição de justiça se manter, a distância dos valores preditos para cada indivíduo deve ser no máximo a distância entre os indivíduos (Dwork et al., 2012).

- **Justiça por não uso**

Um classificador binário satisfaz essa definição se nenhuma variável sensível z for explicitamente usada no processo de tomada de decisão. Ou seja, a variável não será usada no treinamento e a predição deve ser a mesma para indivíduos com o mesmo conjunto de atributos \mathbf{x}_j (Kusner et al., 2017).

- **Justiça por contrafactual**

Essa definição de justiça garante que a predição de um indivíduo seria a mesma caso a variável sensível fosse diferente (Kusner et al., 2017). No caso do COMPAS, a probabilidade de atribuir um risco alto para uma pessoa negra seria a mesma se ela fosse uma pessoa branca.

3.4 Soluções para mitigar o problema de justiça

Além das definições apresentadas na Seção 3.3 com o objetivo de identificar e quantificar o problema de justiça nos modelos de aprendizado de máquina, soluções para mitigar o problema também foram propostas. Como não é objetivo deste trabalho aprofundar ou propor soluções, aqui só serão citadas algumas delas. Em Goel et al. (2018); Kamishima et al. (2012); Krasanakis et al. (2018); Menon & Williamson (2018) foram propostas soluções para satisfazer algumas das definições de justiça. Zafar et al.

(2017) propôs um *framework* para projetar classificadores que não sofram de discriminação por tratamento, impacto e maus-tratos. Em Wu et al. (2018) foi proposto um *framework* que garante que o classificador será justo a partir de restrições impostas nos dados de treinamento.

Também surgiram vários projetos e plataformas que procuram endereçar o problema de justiça de uma maneira mais genérica. A Tabela 3.3 apresenta algumas dessas propostas.

Tabela 3.3. Visão geral dos projetos endereçando o problema de justiça em AM. Fonte: adaptado de Caton & Haas (2020)

Projeto	Descrição
AIF360 (Bellamy et al., 2018)	Ferramenta de código aberto que permite examinar, reportar e mitigar casos de discriminação e vieses em modelos de aprendizado de máquina.
Fairlearn ⁵	Ferramenta que é composta por dois componentes principais: algoritmos para mitigar casos de injustiça e um painel para avaliar os grupos que são negativamente impactados pelo modelos. Também é possível comparar múltiplos modelos em termos de várias métricas e definições de justiça.
Aequitas (Saleiro et al., 2018)	Ferramenta de código aberto para auditar modelos de aprendizado de máquina.
Responsibly ⁶	Framework que disponibiliza conjunto de dados, métricas e algoritmos para medir e mitigar vieses em modelos de classificação e processamento de linguagem natural.
Audit AI ⁷	Ferramenta que mede e mitiga os efeitos de discriminação nos dados de treinamento e predições.
ML Fairness Gym ⁸	Framework para estudar e investigar efeitos a longo prazo de justiça em cenários simulados onde um agente de aprendizado interage com o ambiente ao longo do tempo.

Neste capítulo nós apresentamos o conceito de aprendizado de máquina, os principais algoritmos utilizados nos problemas de classificação e por fim os principais conceitos de justiça presentes na literatura. No Capítulo 4 serão apresentadas as propostas de definições de justiça adaptadas para modelos de análise de sobrevivência, com base no que foi visto nos Capítulos 2 e 3 desta dissertação.

⁵<https://fairlearn.org/>

⁶<https://github.com/ResponsiblyAI/responsibly>

⁷<https://github.com/pymetrics/audit-ai>

⁸<https://github.com/google/ml-fairness-gym>

Capítulo 4

Definições de justiça para modelos de análise de sobrevivência

No Capítulo 2 foi apresentado o modelo de Cox e no Capítulo 3 as principais definições de justiça para modelos de aprendizado de máquina. Neste capítulo iremos mostrar as propostas de definições de justiça adaptadas para modelos de análise de sobrevivência.

4.1 Propostas

A área de justiça em aprendizado de máquina ainda possui muitas oportunidades de pesquisa. As propostas sugeridas neste capítulo são fortalecidas pelo fato de não haver outro trabalho na área que tenha proposto alguma definição para modelos de análise de sobrevivência.

Com isso em mente, seguem 3 abordagens que foram expandidas para quatro propostas de uso em modelos de análise de sobrevivência. No nosso caso, as quatro foram usadas em conjunto com o modelo de Cox e duas também usaram o método de Kaplan-Meier.

A primeira abordagem foca na disparidade entre curvas de sobrevivência observadas nos dados quando comparadas com previsões. Ou seja, a curva de sobrevivência empírica extraída via Kaplan-Meier (Kaplan & Meier, 1958) apresenta uma diferença estatística entre dois grupos, porém a curva de predição do modelo não captura tal diferença. Essa abordagem foi dividida em duas propostas: 1. Divergência em paridade demográfica e 2. Divergência em paridade demográfica condicionada.

A segunda abordagem é uma métrica denominada discriminação causal (Galhotra et al., 2017) que consiste de um cálculo do c-index (Harrell et al., 1982) onde altera-

mos nos dados o grupo de interesse. E por fim, propomos uma métrica nova chamada de justiça de filas, onde comparamos cenários hipotéticos de duas pessoas sendo julgadas por um modelo de aprendizado de máquina no mesmo tempo. Aqui buscamos discrepâncias entre o tempo observado do evento e a chance de sobrevivência.

4.1.1 Proposta 1: Divergência em paridade demográfica / estatística

Nesse primeiro caso usamos as curvas de sobrevivência dos indivíduos, visto na equação 2.12:

$$S(t) = S_0(t)^{\exp(\mathbf{x}\beta)}$$

e adaptamos de acordo com a definição de paridade demográfica, visto na equação 3.2, tal que:

$$(S(t|z = 0) = S(t|z = 1)) \approx (\hat{S}(t|z = 0) = \hat{S}(t|z = 1)) \quad (4.1)$$

onde $S(t)$ é a curva observada, $\hat{S}(t)$ é a curva predita e $z \in \{0, 1\}$ é o atributo sensível. Aqui é importante lembrar que enquanto as definições vistas no Capítulo 3 são probabilidades dadas em relação a classes, na área de análise de sobrevivência temos curvas que representam a probabilidade no tempo. Por isso, nossas definições fazem essa adaptação para considerar somente a curva $\hat{S}(t)$ e não $p(\hat{y})$.

Para obter a curva $S(t)$ usamos o método de Kaplan-Meier enquanto a curva predita $\hat{S}(t)$ é obtida através do modelo de Cox. Se as curvas obtidas empiricamente forem similares as curvas preditas, dizemos que a proposta foi satisfeita e o resultado é considerado justo. Se forem divergentes então a proposta não foi satisfeita e o resultado será considerado injusto.

Um exemplo de uso dessa definição, com dados do MIMIC-III, pode ser vista na Figura 4.1. Nesse caso temos dois gráficos, o da esquerda com as curvas empíricas e o da direita com as curvas preditas de sobrevivência, em ambos os casos uma curva é de pessoas negras e outra de pessoas brancas. As linhas pontilhadas representam a mediana de cada curva e o sombreado representa o intervalo de confiança de 95%. Neste caso as curvas $S(t)$ e $\hat{S}(t)$ não apresentaram divergências e por isso seu resultado é considerado justo. Outros casos e a análise aprofundada serão mostrados no Capítulo 6.

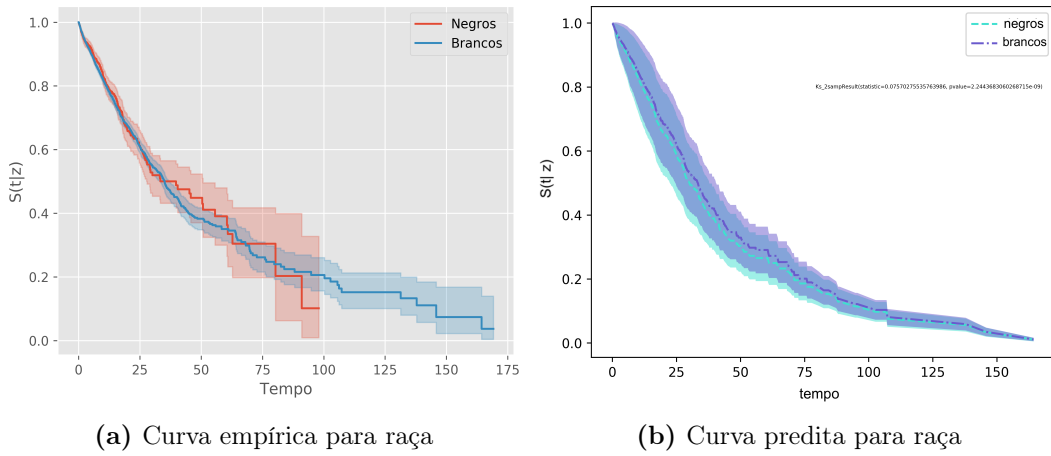


Figura 4.1. Divergência em paridade demográfica para modelos de análise de sobrevivência

4.1.2 Proposta 2: Divergência em paridade demográfica condicionada

Nesse segundo caso também usamos as curvas de sobrevivência dos indivíduos e adaptamos para a definição divergência em paridade demográfica condicionada, vista na equação 3.3, tal que:

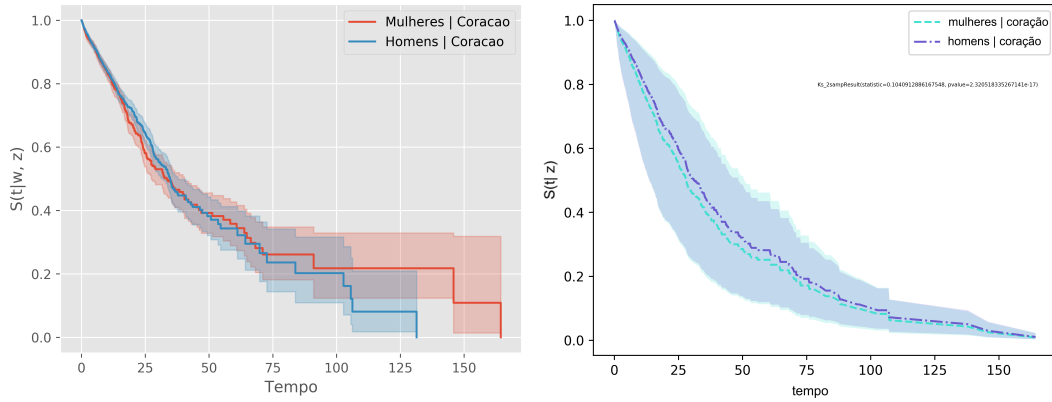
$$S(t|\mathbf{w}, z = 0) = S(t|\mathbf{w}, z = 1) \approx \hat{S}(t|\mathbf{w}, z = 0) = \hat{S}(t|\mathbf{w}, z = 1), \quad (4.2)$$

onde $S(t)$ é a curva empírica, $\hat{S}(t)$ é a curva predita, \mathbf{w} são atributos que influenciam a resposta e $z \in \{0, 1\}$ é o atributo sensível. As curvas são obtidas da mesma maneira que na proposta 1, através do método Kaplan-Meier e modelo de Cox.

Um exemplo de uso dessa definição, também com dados do MIMIC-III, pode ser vista na Figura 4.2. Nesse caso temos dois gráficos, o da esquerda com os dados empíricos e o da direita com a predição, em ambos há uma curva para mulheres e outra para homens e o atributo escolhido foi problemas no coração. Novamente, as linhas pontilhadas representam a mediana de cada curva e o sombreado representa o intervalo de confiança de 95%. Outros casos e a análise aprofundada serão mostrados no Capítulo 6. Aqui, as curvas $S(t)$ e $\hat{S}(t)$ não apresentaram divergências e por isso seu resultado é considerado justo.

4.1.3 Proposta 3: Discriminação causal

A terceira proposta é a utilização do erro dado pelo c-index para comparação dos resultados entre indivíduos similares que diferem apenas na variável sensível z . Relem-



(a) Curva empírica para gênero | coração (b) Curva predita para gênero | coração

Figura 4.2. Divergência em paridade demográfica condicionada para modelos de análise de sobrevivência

brando a definição apresentada no Capítulo 3, temos que o resultado produzido deve ser o mesmo para quaisquer dois indivíduos a e b que possuam um conjunto de atributos \mathbf{x}_j muito parecidos entre si mas que diferem na variável sensível z . Adicionalmente, temos a definição de c -index vista em 2.15:

$$C_{td} = \frac{\sum_{i,j} 1_{\hat{S}(t_i|\mathbf{x}_i) < \hat{S}(t_j|\mathbf{x}_j)} \cdot 1_{t_i < t_j} \cdot \delta_i}{\sum_{i,j} 1_{t_i < t_j} \cdot \delta_i}$$

tal que o erro é definido como:

$$erro(z) = 1 - C_{td}(z) \tag{4.3}$$

A ideia é que $erro(z_1) = erro(z_2)$, onde z_1, z_2 são grupos com variáveis sensíveis diferentes entre si. Como as bases utilizadas apresentaram dificuldades em achar pares com características similares, fizemos o arranjo entre todos os pares disponíveis. Exemplos de uso dessa proposta e a análise dos casos serão mostrados nos Capítulos 5 e 6.

4.1.4 Proposta 4: Justiça de filas

A quarta proposta envolve novamente erro porém com uma versão modificada do c -index para permitir comparação entre grupos com diferentes variáveis sensíveis, onde a modificação sugerida permite que a comparação seja feita somente entre indivíduos de grupos diferentes e não entre indivíduos aleatórios como na definição original.

A proposta de um nova métrica de risco vai ao encontro do que é sugerido na medicina, onde escores de risco auxiliam prestadores da área da saúde a melhorar o suporte ao paciente. No entanto, o desenvolvimento desses escores não é trivial nem possui um processo definido (Pirracchio et al., 2015; Xie et al., 2020). A nossa proposta para essa métrica é definida como:

$$C_{td}(z_1, z_2) = \frac{\sum_{i \in z_1, j \in z_2} 1_{\hat{S}(t_i|\mathbf{x}_i) < \hat{S}(t_j|\mathbf{x}_j)} \cdot 1_{t_i < t_j} \cdot \delta_i}{\sum_{i,j} 1_{t_i < t_j} \cdot \delta_i} \quad (4.4)$$

tal que o erro:

$$erro(z_1, z_2) = 1 - C_{td}(z_1, z_2) \quad (4.5)$$

onde z_1, z_2 são grupos com variáveis sensíveis diferentes entre si. A ideia dessa métrica é verificar se a diferença entre os grupos resulta em uma situação de injustiça.

Por exemplo, imagine que o grupo z_1 representa pessoas negras e o grupo z_2 representa pessoas brancas. Além disso, considere que duas pessoas, uma negra e uma branca cheguem para atendimento ao mesmo tempo em uma UTI. A métrica acima vai capturar a fração de vezes, em um conjunto de treino, que o grupo z_1 teve um tempo de falecimento menor que o grupo z_2 porém a previsão de sobrevivência ($\hat{S}(\cdot)$) fez a previsão oposta. Isto é, que a pessoa negra teve uma previsão de tempo de mortalidade maior. Por fim, no caso das duas pessoas chegarem na UTI no mesmo tempo, isto indica uma chance de erro na priorização do tratamento.

A Figura 4.3 ilustra um exemplo, onde os círculos pretos representam as curvas de sobrevivência dos indivíduos não censurados e $S1_{z1} < S2_{z1} < S3_{z2} < S4_{z2} < S5_{z1}$. Portanto $S1_{z1}$ é comparado somente com $S3_{z2}$ e $S4_{z2}$, enquanto o $S3_{z12}$ é comparado com o $S5_{z1}$.

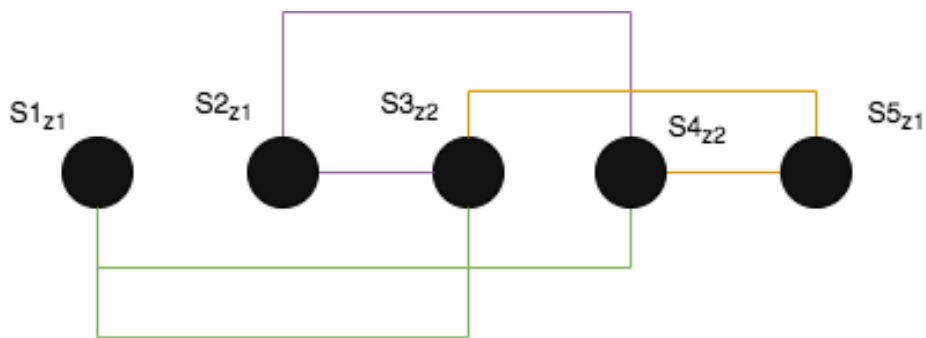


Figura 4.3. Cálculo da métrica justiça de filas

Exemplos de uso dessa proposta e a análise das comparações entre os diferentes

grupos será aprofundada nos Capítulos 5 e 6.

Neste capítulo apresentamos três propostas de adaptação das definições de justiça para modelos de análise de sobrevivência e uma proposta de modificação do c-index para verificar a existência de injustiça em casos de comparação de grupos. No Capítulo 5 serão apresentadas as bases de dados, as análises feitas com esses dados e por fim a parte de experimentação. No Capítulo 6 teremos as aplicações das propostas vistas neste capítulo no modelo escolhido e as discussões envolvendo justiça.

Capítulo 5

Experimentos e resultados

Neste capítulo iremos mostrar os experimentos e resultados decorrentes da pesquisa em análise de sobrevivência e justiça em aprendizado de máquina. Em particular iremos começar mostrando as bases de dados escolhidas e a motivação para escolha das mesmas. Em seguida serão apresentados as análises e os experimentos feitos e por fim quais foram os resultados provenientes do modelo escolhido. As discussões envolvendo as métricas de justiça ocorrerão no Capítulo 6.

5.1 Bases de dados

Para essa dissertação foram escolhidas três bases de dados: MIMIC-III, Rossi e COMPAS. A primeira é uma base médica com dados de um hospital nos EUA, a segunda e terceira são da área criminal.

Quando falamos de modelos de análise de sobrevivência, os principais casos de uso são com dados médicos, geralmente para modelar o tempo até a morte de um paciente. Além disso, os diversos avanços em inteligência artificial (ex. DeepMind, Siri, carros autônomos) também estão transformando a área da saúde. Existe uma definição denominada *grupos protegidos*, que se refere a populações que já sofreram / ainda sofrem de experiências preconceituosas e que portanto estão ainda mais vulneráveis a possíveis erros por parte dos modelos preditivos (Rajkomar et al., 2018). Por esse motivo Rajkomar et al. (2018), Goodman et al. (2018) e Char et al. (2018) salientam que é necessário fortalecer a ideia de assegurar justiça nos modelos que serão usados na área de saúde, dado que o mais importante é que todos os pacientes sejam beneficiados.

Na área criminal, instrumentos de análise de risco, incluindo modelos preditivos, estão ganhando popularidade no sistema judiciário dos EUA e sendo usados para decisões que envolvem prisão preventiva, liberdade condicional e condenações (Choul-

dechova, 2017). Esses são casos de uso que afetam diretamente vidas humanas e trazem a tona as preocupações de justiça resultante das decisões tomadas com base nesses instrumentos.

- **MIMIC-III**

MIMIC-III, sigla para *Medical Information Mart for Intensive Care* versão III, é uma base médica com dados deidentificados de pacientes que foram internados no hospital Beth Israel Deaconess Medical Center em Boston, Massachusetts - EUA entre os anos de 2001 e 2012. Essa base de dados foi criada com o intuito de disponibilizar gratuitamente informação para a comunidade científica e outros, dado a falta de dados disponíveis para pesquisa. Ela foi desenvolvida pela equipe do MIT Lab for Computational Physiology e possui dados anônimos associados com aproximadamente 40000 pacientes da unidade de tratamento intensivo (UTI). Ela inclui dados demográficos, sinais vitais, testes de laboratório, medicamentos e outros (Johnson et al., 2016). Apesar da base ser gratuita, por conter informações sensíveis e detalhadas de milhares de pacientes, é necessária uma solicitação especial, treinamento para uso de dados humanos em pesquisa e adesão aos termos da *data use agreement* (DUA) para finalmente ter acesso a base completa.

- **Rossi**

Essa base contém dados de um experimento de reincidência feito com 432 presos condenados que foram soltos das prisões estaduais do estado de Maryland - EUA na década de 1970 (Rossi et al., 1980). Eles foram acompanhados por um ano após serem soltos e o experimento consistia em escolher aleatoriamente as pessoas, dar um auxílio financeiro para uma metade enquanto a outra metade não recebia nada e avaliar o quanto isso afetava a reincidência. A base usada foi obtida através da biblioteca Lifelines¹. A base Rossi é uma base clássica na análise de sobrevivência e o foco da mesma é a previsão de reincidência criminal. Por ser uma base clássica e pequena estudaremos ela com modelos de sobrevivência já consolidados.

- **COMPAS**

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) é uma ferramenta desenvolvida pela Northpointe para prever a chance de uma pessoa que foi presa em reincidir, ou seja, repetir um crime ou delito. A base

¹<https://github.com/CamDavidsonPilon/lifelines>

utilizada foi obtida através de requisição pública e contém dois anos de registros do condado de Broward na Florida / EUA. Os dados estão disponíveis no repositório da ProPublica². Em particular, fazemos uso do arquivo *cox-parsed.csv* no qual é possível fazer uma análise similar à da base Rossi, isto é, previsão da reincidência criminal como sendo o evento de interesse.

5.2 Análises exploratórias

Nessa seção vamos trazer as análises que foram feitas em cada uma dessas bases, mostrando quais dados foram utilizados, o que eles significam, a análise exploratória e os diferentes algoritmos testados para a análise de sobrevivência.

5.2.1 MIMIC-III

O MIMIC-III é uma base de dados relacional contendo tabelas com dados de pacientes da unidade de tratamento intensivo do hospital Beth Israel Deaconess Medical Center. Cada tabela consiste de linhas e colunas, onde cada coluna contém um atributo (ex. identificador do paciente) e cada linha contém um valor atrelado ao atributo (ex. 33).

A base possui uma série de tabelas, porém nesta dissertação usamos quatro delas:

- ICUSTAY DETAIL: Todas as estadias na UTI contendo os dados demográficos dos pacientes.
- DIAGNOSES ICD: Diagnósticos existentes no hospital, codificados usando o sistema de Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde (CID)
- DESCRIPTION DIAGNOSES ICD: A descrição de cada diagnóstico
- OASIS: Sigla para *Oxford Acute Severity of Illness Score* (Johnson et al., 2013), é uma pontuação que indica a gravidade da doença do paciente que está na UTI. Essa tabela contém a pontuação para cada internação no hospital.

As tabelas são conectadas por identificadores que possuem o sufixo 'ID'. Logo cada ICUSTAY_ID corresponde a um único HADM_ID e um único SUBJECT_ID. Cada HADM_ID corresponde a um único SUBJECT_ID. Um SUBJECT_ID pode ter

²<https://github.com/propublica/compas-analysis>

múltiplos HADM_ID (múltiplas internações de um mesmo paciente) e múltiplos ICUSTAY_ID (múltiplas entradas na UTI, podendo ser dentro da mesma hospitalização ou em múltiplas hospitalizações).

Como as tabelas representam os dados de aproximadamente 40000 pacientes, optamos por selecionar casos entre quatro tipos de diagnósticos: transplantes, câncer, diabetes e coração. Que estão entre as principais causas de morte em países de alta renda³, lembrando que estamos trabalhando com um base com dados dos EUA.

Nesta primeira etapa de limpeza e tratamento dos dados foram usados os pacotes pandas⁴, numpy⁵, psycopg2⁶, matplotlib⁷, seaborn⁸ e tableone⁹, em conjunto com Jupyter Notebook e a linguagem Python.

Após unir e fazer todos os tratamentos necessários nos dados selecionados, criamos uma única tabela com 9101 linhas. As colunas booleanas foram criadas para identificar quais doenças os pacientes tinham e facilitar posteriormente o uso da regressão. Os atributos estão descritos na Tabela 5.1.

³<https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>

⁴<https://pandas.pydata.org/>

⁵<https://numpy.org/>

⁶<https://www.psycopg.org/>

⁷<https://matplotlib.org/>

⁸<https://seaborn.pydata.org/>

⁹<https://github.com/tompollard/tableone>

Tabela 5.1. Atributos presentes na tabela final usada para os experimentos.

Atributo	Descrição
subject_id	identificador do paciente
hadm_id	identificador da hospitalização
icustay_id	identificador da internação na UTI
admission_type	tipo de admissão no hospital
age	idade do paciente
ethnicity_grouped	raça do paciente
gender	gênero do paciente
insurance	seguro de saúde do paciente
hospital_expire_flag	se o paciente morreu durante a hospitalização
oasis_score	pontuação da gravidade da doença do paciente
icd9_code	identificador do diagnóstico do paciente dado pelo hospital
los_hospital	duração da hospitalização do paciente, em dias
icd_transplant	booleano indicando se foi caso de transplante
icd_cancer	booleano indicando se foi caso de câncer
icd_diabetes	booleano indicando se foi caso de diabetes
icd_heart	booleano indicando se foi caso de coração
icd_alzheimer	booleano indicando se foi caso de Alzheimer

Em seguida foi feita uma análise exploratória para entender a distribuição dos dados de acordo com os atributos selecionados. Na Figura 5.1 é possível visualizar um sumário dos pacientes considerando gênero, raça, idade, seguro saúde e duração da hospitalização.

No sumário da Figura 5.1 vemos que a grande maioria dos pacientes são pessoas brancas que usam o sistema público de saúde. 44% dos pacientes são mulheres, a mediana de dias hospitalizado é 8, e a mediana da idade fica entre 63 e 71 anos.

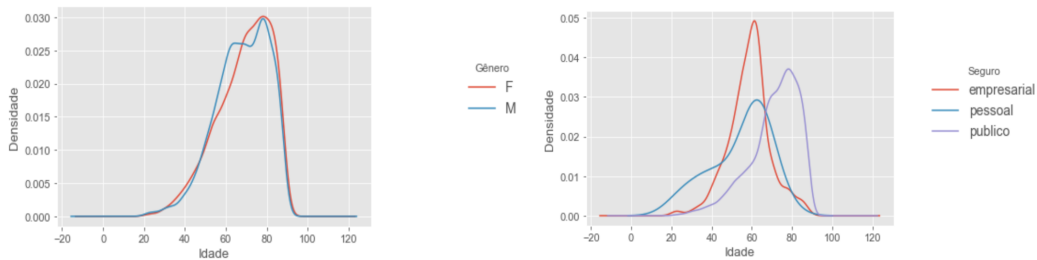
Nas Figuras 5.2, 5.3 e 5.4 continuamos a análise para entender como é a distribuição dos valores para diferentes recortes. Em um primeiro momento olhamos para a distribuição do gênero (somente feminino e masculino) e seguro saúde considerando a idade. Para o gráfico de gênero, ambos possuem uma média de idade acima dos 60 anos. Já para o gráfico de seguro, a média de idade diverge um pouco mais, sendo o público usado por pessoas mais idosas, já na faixa dos 70-80 anos enquanto os outros dois são usados mais por pessoas na faixa dos 60 anos.

Em seguida, é apresentada a distribuição do tempo que as pessoas ficam hospitali-

	Overall	asiático	branco	hispanico	negro	
n	9101	267	7363	314	1157	
Gênero, n (%)	F	3965 (43.6)	98 (36.7)	3092 (42.0)	138 (43.9)	637 (55.1)
	M	5136 (56.4)	169 (63.3)	4271 (58.0)	176 (56.1)	520 (44.9)
Seguro saúde, n (%)	empresarial	2041 (22.4)	45 (16.9)	1782 (24.2)	55 (17.5)	159 (13.7)
	pessoal	17 (0.2)		10 (0.1)	2 (0.6)	5 (0.4)
	publico	7043 (77.4)	222 (83.1)	5571 (75.7)	257 (81.8)	993 (85.8)
Duração da hospitalização, median [Q1,Q3]	7.9 [4.6,14.0]	7.7 [3.9,14.4]	8.0 [4.7,14.0]	7.1 [4.1,14.2]	7.9 [4.3,14.1]	
Idade, median [Q1,Q3]	69.9 [59.8,78.8]	67.2 [57.4,77.7]	70.9 [60.7,79.5]	63.0 [51.2,72.4]	67.0 [57.1,75.7]	

Figura 5.1. Sumário estatístico para a população de pacientes selecionados

zadas recebendo algum tipo de tratamento. A grande maioria dos pacientes permanece poucos dias e alguns poucos casos se estende até 6 meses de internação. Por fim, a Figura 5.4 mostra a distribuição do Oasis considerando gênero e raça, e que não apresenta diferença significativa entre os grupos.



(a) Distribuição do gênero por idade (b) Distribuição do seguro de saúde por idade

Figura 5.2. Gráficos de densidade para gênero e seguro de saúde

No Capítulo 6, iremos apresentar e discutir os resultados das curvas de sobrevivência empíricas e preditas para os vários grupos estudados.

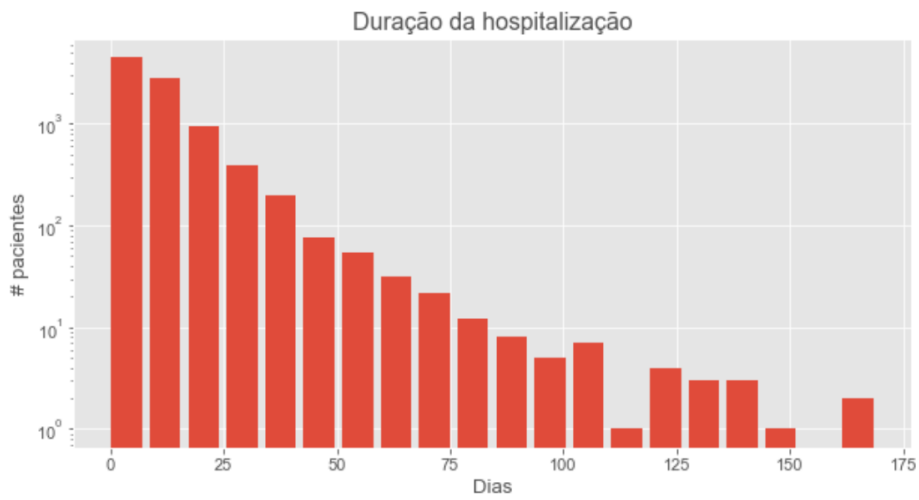
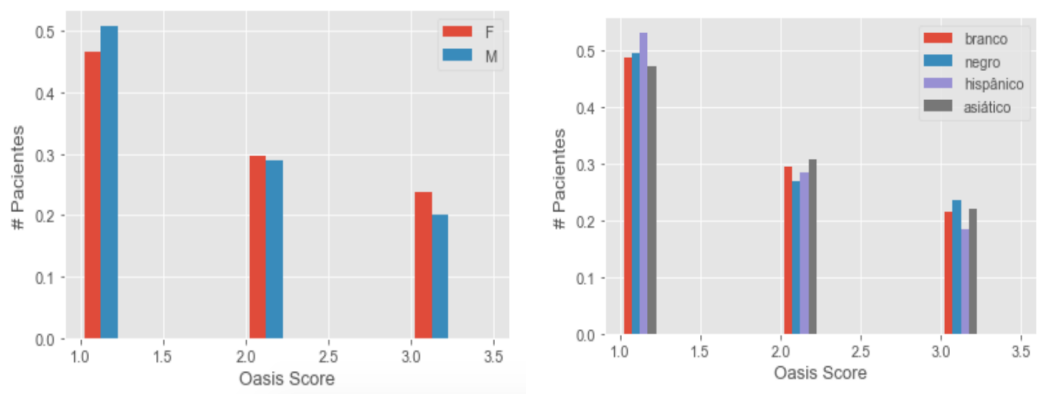


Figura 5.3. Gráfico com distribuição dos tempos de hospitalização



(a) Gráfico com distribuição da pontuação Oasis para cada gênero (b) Gráfico com distribuição da pontuação Oasis para cada raça

Figura 5.4. Pontuação Oasis para gênero e raça

5.2.2 Rossi

A base de dados Rossi possui 432 linhas e é resultado de um estudo randomizado controlado cuja a variável controlada foi a do financiamento. A Tabela 5.2 apresenta os atributos da mesma.

Tabela 5.2. Atributos presentes na tabela final usada para os experimentos.

Atributo	Descrição
week	semana da primeira prisão após soltura.
arrest	booleana que indica se foi preso durante o período do estudo (1 ano). 1 para sim, 0 para não.
fin	booleana que indica se a pessoa recebeu o auxílio financeiro ao ser solto da prisão. 1 para sim, 0 para não.
age	idade ao ser solto.
race	raça. 1 para negros, 0 para outros.
wexp	booleana para indicar se a pessoa tinha experiência no mercado de trabalho antes de ser presa. 1 para sim, 0 para não.
mar	booleana para indicar se a pessoa era casada ao ser solta. 1 para sim, 0 para não.
paro	booleana para indicar se a pessoa foi solta em liberdade condicional. 1 para sim, 0 para não.
prio	número de condenações anteriores.

No sumário que vemos na Figura 5.5 a grande maioria dos presos eram pessoas negras, que não eram casadas e foram soltas em forma de liberdade condicional. Em seguida, selecionamos 25 indivíduos aleatoriamente para analisar o gráfico de censura, presente na Figura 5.6. Nesse gráfico há indivíduos que reincidiram durante o acompanhamento de 52 semanas e são representados pelas linhas vermelhas e indivíduos que não reincidiram e são representados pelas linhas azuis. Por fim, na Figura 5.7, olhamos a distribuição de idade e prisões anteriores em relação a raça. Como negros representam a maior parte da base isso é refletido em ambos os gráficos.

		Overall	0	1
	n	432	53	379
arrest, n (%)	0	318 (73.6)	41 (77.4)	277 (73.1)
	1	114 (26.4)	12 (22.6)	102 (26.9)
fin, n (%)	0	216 (50.0)	31 (58.5)	185 (48.8)
	1	216 (50.0)	22 (41.5)	194 (51.2)
wexp, n (%)	0	185 (42.8)	20 (37.7)	165 (43.5)
	1	247 (57.2)	33 (62.3)	214 (56.5)
mar, n (%)	0	379 (87.7)	44 (83.0)	335 (88.4)
	1	53 (12.3)	9 (17.0)	44 (11.6)
paro, n (%)	0	165 (38.2)	23 (43.4)	142 (37.5)
	1	267 (61.8)	30 (56.6)	237 (62.5)

Figura 5.5. Sumário estatístico para a base Rossi

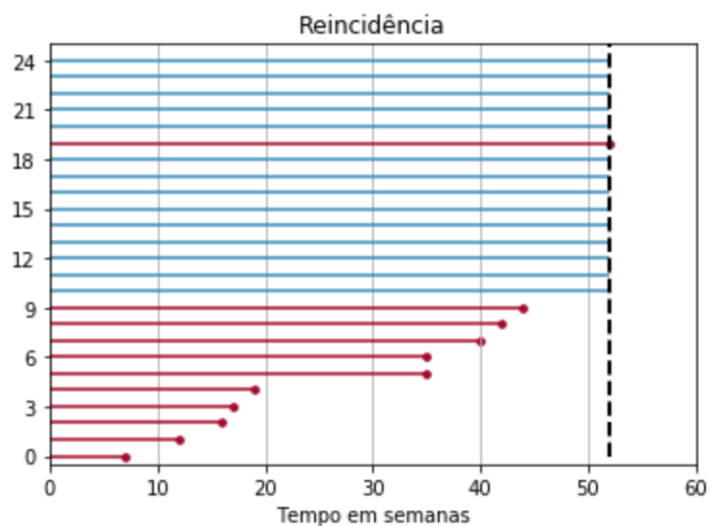
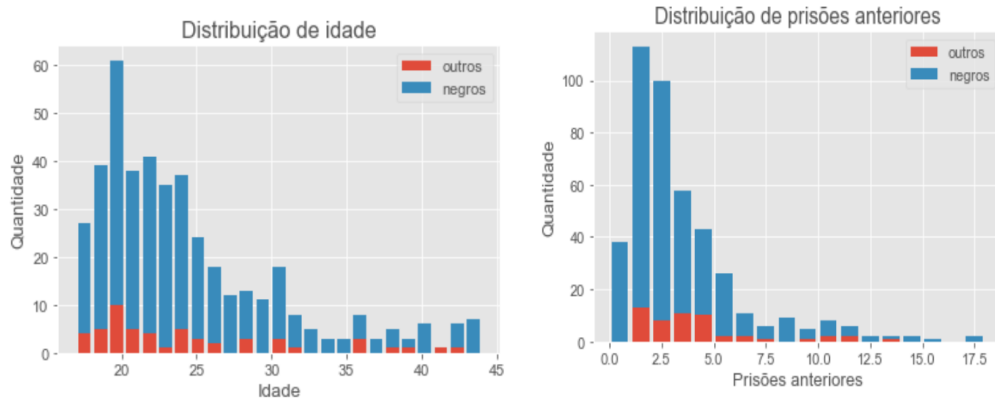


Figura 5.6. Gráfico de censura para 25 indivíduos da base Rossi.



(a) Distribuição da idade para o atributo *race* (b) Distribuição do número de prisões anteriores para o atributo *race*

Figura 5.7. Análise das distribuições de idade e prisões anteriores

5.2.3 COMPAS

Essa base possui inicialmente 13419 registros e 52 colunas. Após um tratamento nos dados permaneceram 10314 registros e o número de colunas aumentou para 54. Entre os atributos presentes na base estão: nome, gênero, idade, raça, quantidade de prisões anteriores, reincidência, tipo de acusação e fator de risco.

No sumário estatístico presente na Figura 5.8 a grande maioria dos presos se dividem entre negros e brancos, do gênero masculino, com baixo fator de risco. No entanto, negros representam 73% do grupo considerado de alto risco. Valor muito maior do que qualquer outra raça presente nessa mesma base.

	Overall	0_Caucasian	African-American	Asian	Hispanic	Native American	Other	
n	10314	3569	5147	51	944	32	571	
sex, n (%)	Female	2112 (20.5)	862 (24.2)	973 (18.9)	5 (9.8)	158 (16.7)	7 (21.9)	107 (18.7)
	Male	8202 (79.5)	2707 (75.8)	4174 (81.1)	46 (90.2)	786 (83.3)	25 (78.1)	464 (81.3)
score_text, n (%)	factor_a_low	5751 (55.8)	2372 (66.5)	2184 (42.4)	38 (74.5)	684 (72.5)	16 (50.0)	457 (80.0)
	factor_b_medium	2611 (25.3)	800 (22.4)	1543 (30.0)	9 (17.6)	168 (17.8)	9 (28.1)	82 (14.4)
	factor_c_high	1952 (18.9)	397 (11.1)	1420 (27.6)	4 (7.8)	92 (9.7)	7 (21.9)	32 (5.6)

Figura 5.8. Sumário estatístico para a base COMPAS

Para essa base também selecionamos aleatoriamente 25 indivíduos para analisar o gráfico de censura, presente na Figura 5.9. Em seguida, na Figura 5.10 são apresentadas as distribuições de idade, prisões anteriores e fator de risco em função da raça. No gráfico (a) da Figura 5.10 observa-se que a média de idade varia entre 20-40 anos.

O gráfico (b) mostra que a maior parte das pessoas não possui prisões anteriores. No gráfico (c) há uma quantidade maior de pessoas consideradas de risco baixo. É possível observar que negros e brancos representam a maior parte da base.

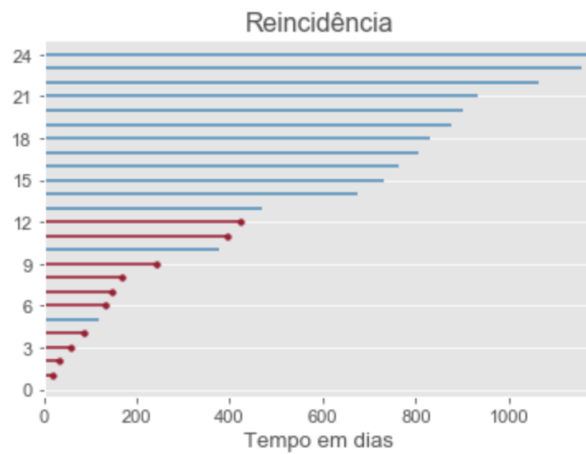


Figura 5.9. Gráfico de censura para 25 indivíduos da base COMPAS.

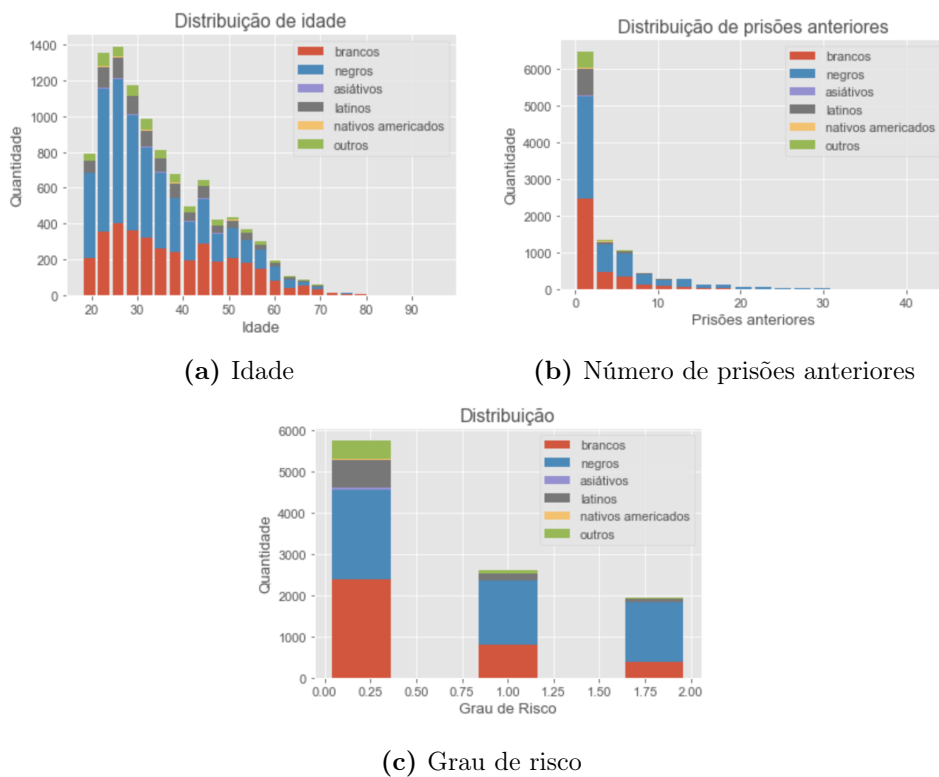


Figura 5.10. Análise das distribuições do atributo *race* para base COMPAS

5.3 Modelos

Após as etapas de tratamento e análise exploratória, seguimos para a modelagem, onde fizemos o treinamento dos algoritmos e otimização dos parâmetros. Na próxima seção iremos avaliar as métricas de sucesso e apresentar os resultados do modelo escolhido.

Nessa fase foram usadas as seguintes bibliotecas: Lifelines, Scikit-Survival¹⁰, Pycox¹¹, PyTorch¹², Hyperopt¹³. As três primeiras são bibliotecas que possuem algoritmos de análise de sobrevivência, PyTorch é uma biblioteca para aprendizado profundo e o Hyperopt é uma biblioteca para otimização dos parâmetros dos modelos.

Todos os notebooks e scripts criados durante o desenvolvimento desta dissertação estão disponíveis no Github¹⁴ e podem ser usados de acordo com a licença GNU General Public License v3.0.

5.3.1 MIMIC-III

Na etapa de modelagem, foram treinados e otimizados sete algoritmos de análise de sobrevivência. Todos eles foram testados com cinquenta sementes diferentes, que foram geradas aleatoriamente a partir da função `os.urandom()` do python. Para essa base, a coluna `hospital_expire_flag` representa o evento de interesse e a coluna `los_hospital` representa o tempo do evento de interesse.

Um ponto importante é que iniciamos a modelagem com 9101 hospitalizações que representavam 6379 pacientes distintos. Porém ao analisar mais profundamente, há pessoas com um grande número de hospitalizações e por isso, optamos por usar somente dados de pessoas com até 3 hospitalizações, isso reduziu o número de amostras para 7715 hospitalizações e 6116 pacientes distintos. Essa decisão se deu para balancear a quantidade de hospitalizações entre os pacientes.

5.3.1.1 Algoritmos

Segue abaixo os algoritmos utilizados para essa base com uma breve explicação sobre cada um:

- **CoxPHFitter**

¹⁰<https://github.com/sebp/scikit-survival>

¹¹<https://github.com/havakv/pycox>

¹²<https://pytorch.org/>

¹³<https://github.com/hyperopt/hyperopt>

¹⁴<https://github.com/bseewald/fairness-ml-health>

Algoritmo presente na biblioteca Lifelines. É a implementação do modelo de Cox de riscos proporcionais, visto no Capítulo 2. Para treinamento a base foi dividida entre treino, validação e teste, numa proporção de 50 / 12,4 / 37,6 respectivamente.

Essa divisão é importante pois a função preditora depende diretamente dos parâmetros de entrada e ao treinar e testar sobre os mesmos dados estamos cometendo um grave erro metodológico. Isso pode levar a situação de *overfitting*, que é quando o modelo acerta tudo no treinamento porém tem uma alta taxa de erro ao classificar dados nunca vistos antes. Por isso é uma prática comum avaliar o modelo usando dados de validação e teste (Pedregosa et al., 2011).

O gráfico da Figura 5.11 mostra como o erro de predição evolui em ambas as amostras de treino e teste considerando a complexidade do modelo. É possível identificar que o erro é alto quando a complexidade é baixa e volta a ficar alta quando a complexidade é alta. O ponto ideal acaba sendo uma ponderação do erro e da complexidade do modelo.

Ainda sim, há a possibilidade de haver *overfitting* na amostra de teste. Por isso também é sugerido aplicar o método de validação cruzada, que nada mais é do que dividir a amostra de treinamento em k pedaços ("folds") e usar $k - 1$ amostras para treinar. O pedaço restante é usado para avaliar a performance do modelo. Esse processo ocorre k vezes. A Figura 5.12 ilustra a validação cruzada.

O passo seguinte é a otimização dos parâmetros. Essa técnica visa encontrar um conjunto ótimo de hiperparâmetros dentro de um espaço de busca. Para esse algoritmo foram selecionados os parâmetros *penalizer* e *l1_ratio*. O primeiro é

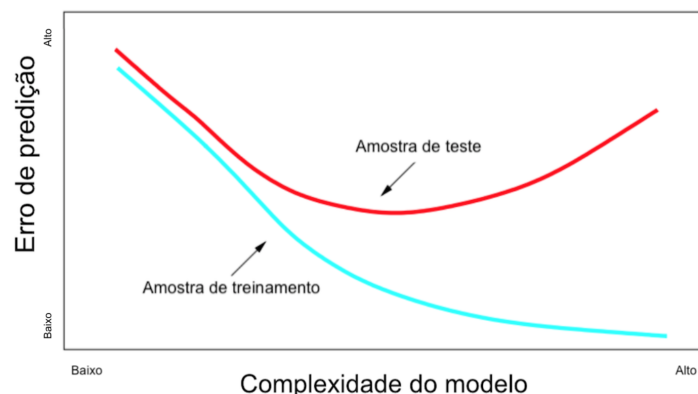


Figura 5.11. Gráfico do erro no treinamento e teste em função da complexidade do modelo. Adaptado de Friedman et al. (2001)

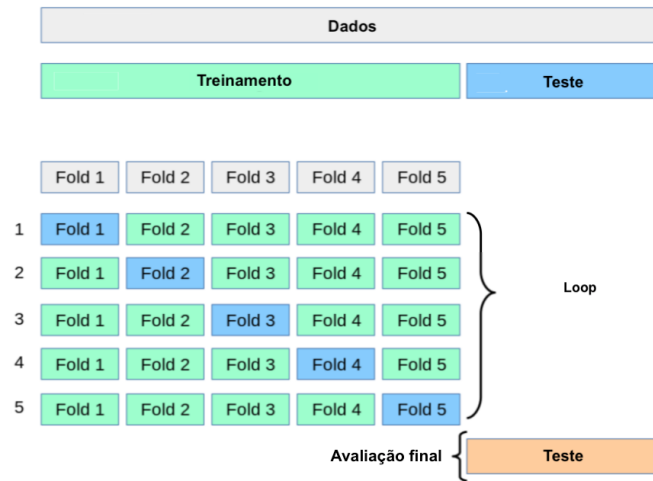


Figura 5.12. Validação cruzada. Adaptado de Pedregosa et al. (2011)

uma penalização que ocorre nos coeficientes durante a regressão. Isso aumenta a estabilidade dos estimadores e controla a alta correlação entre as covariáveis. O segundo é a razão que será usada para a regularização L1 (Lasso) em relação a L2 (Ridge). O objetivo da regularização, em ambos os casos, é que o modelo generalize e não haja *overfitting*. O espaço de busca testado para esse algoritmo está presente na tabela 5.3.

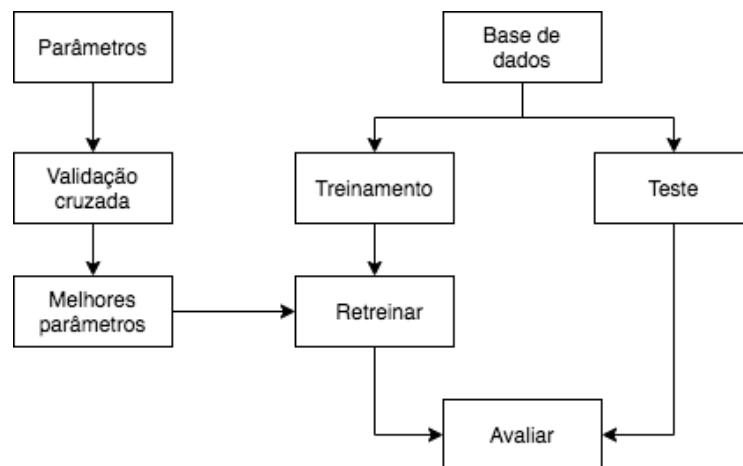


Figura 5.13. Diagrama que mostra o fluxo da modelagem usando validação cruzada e otimização de parâmetros. Adaptado de Pedregosa et al. (2011)

Tabela 5.3. Espaço de busca para os hiperparâmetros

Hiperparâmetro	Valores
penalizer	{100, 10, 1, 0.1, 0.01, 1e-03, 1e-04, 1e-05}
l1_ratio	{0, 0.001, 0.01, 0.1, 0.5, 1}

- **CoxNet**

Algoritmo presente na biblioteca Scikit-Survival. É uma outra implementação do modelo de Cox de riscos proporcionais. Para treinamento a base foi dividida da mesma maneira que o método anterior: treino, validação e teste, numa proporção de 50 / 12,4 / 37,6 respectivamente.

Para esse algoritmo foram selecionados os parâmetros α (que possui o mesmo significado que o *penalizer*) e *l1_ratio*. O espaço de busca testado para esse algoritmo está presente na tabela 5.4.

Tabela 5.4. Espaço de busca para os hiperparâmetros

Hiperparâmetro	Valores
alpha	{100, 10, 1, 0.1, 0.01, 1e-03, 1e-04, 1e-05}
l1_ratio	{0, 0.001, 0.01, 0.1, 0.5, 1}

- **Random Survival Forest (RSF)**

O RSF, proposto por (Ishwaran et al., 2008), está presente na biblioteca Scikit-Survival. É uma implementação diferente das anteriores pois usa uma técnica de combinação de árvores em vez do modelo semi-paramétrico. Essa escolha se deu devido a outros estudos (Wang et al., 2019; Kvamme et al., 2019) que também apresentam esse método como um possível caminho para a análise de sobrevivência.

Para treinamento, a divisão da base seguiu a mesma regra dos algoritmos anteriores. Em relação a otimização, foram selecionados os parâmetros $n_estimators$ - número de árvores, $min_samples_split$ - número mínimo de amostras para dividir um nodo interno e o $min_samples_leaf$ - número mínimo de amostras que deve existir em cada nodo folha. O espaço de busca testado para esse algoritmo está presente na tabela 5.5.

Tabela 5.5. Espaço de busca para os hiperparâmetros

Hiperparâmetro	Valores
n_estimators	{500, 1000}
min_samples_split	{2, 4, 6, 8}
min_samples_leaf	{2, 8, 32, 64, 128}

- **Cox-MLP**

Esse algoritmo está presente na biblioteca Pycox, que é uma biblioteca para análise de sobrevivência usando redes neurais. O Cox-MLP portanto é a implementação do modelo de Cox de riscos proporcionais para aprendizado profundo.

Tanto para o Cox-MLP, quanto para os próximos algoritmos a serem apresentados, há uma diferença na forma como as variáveis categóricas foram codificadas: para redes neurais foram usadas *entity embeddings* (Guo & Berkahn, 2016), enquanto para os modelos de Cox clássicos e RSF foram usados *One Hot (dummy variables)*.

Para todos os algoritmos da biblioteca Pycox, usamos redes multilayer perceptron em que cada camada possui o mesmo número de nós, ativações ReLU e normalização batch entre camadas. Também foi usado *dropout*, *decoupled weight decay* (Loshchilov & Hutter, 2017) e *early stopping*. Para o gradiente descendente estocástico foi usado AdamWR (Loshchilov & Hutter, 2017).

Considerando a grande quantidade de parâmetros que podem ser otimizados em modelos de aprendizado profundo o espaço de busca testado está presente na tabela 5.6.

Tabela 5.6. Espaço de busca para os hiperparâmetros

Hiperparâmetro	Valores
Batch size	{64, 128, 256, 512, 1024}
Dropout	{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7}
Learning Rate	{0.01, 0.001, 0.0001}
Layer	{2, 4}
Nodes per layer	{64, 128, 256, 512}
Penalizer	{0.1, 0.01, 0.001, 0}
Weight decay	{0.4, 0.2, 0.1, 0.05, 0.02, 0.01, 0}

Segue uma rápida explicação de cada um:

Batch size - número de exemplos de treinamento usados em uma iteração

Dropout - usado para regularização

Learning rate - tamanho do passo a ser dado a cada iteração

Layer - número de camadas da rede

Nodes per layer - quantidade de nós por camada

Penalizer - usado para regularização

Weight decay - usado para regularização

- **Cox-PH / DeepSurv**

Algoritmo também presente na biblioteca Pycox e proposto por Katzman et al. (2018). É uma outra implementação do modelo de Cox de riscos proporcionais para aprendizado profundo e a principal diferença para o Cox-MLP é em relação a função de custo (Kvamme et al., 2019).

- **Cox-Time**

Algoritmo também presente na biblioteca Pycox. Visto rapidamente no Capítulo 2, essa é a implementação do modelo de Cox de risco não proporcional. Essa versão é uma forma de contornar as restrições impostas pela necessidade de manter a taxa de risco constante para todas as instâncias.

A sugestão, até então, era agrupar os dados e usar uma versão estratificada do modelo de Cox (Klein & Moeschberger, 2006). A proposta do Cox-Time é usar uma abordagem paramétrica que não exige essa estratificação dos dados (Kvamme et al., 2019).

- **DeepHit**

DeepHit, proposto por Lee et al. (2018), é um algoritmo presente na biblioteca Pycox. É uma abordagem que permite que o relacionamento entre as covariáveis e os riscos mudem no tempo. DeepHit também permite lidar com situações que tenham uma única causa ou múltiplas causas (mais de um evento de interesse).

Além dos hiperparâmetros vistos na tabela 5.6 (com exceção do *penalizer*), o algoritmo permite otimizar mais alguns parâmetros. *Alpha* e *sigma* são constantes usadas na função de perda e *num. durations* é o número de pontos de tempo discretos.

Tabela 5.7. Hiperparâmetros adicionais no DeepHit

Hiperparâmetro	Valores
Alpha	[0 ,1]
Sigma	{0.1, 0.25, 0.5, 1, 2.5, 5, 10, 100}
Num. Durations	{50, 100, 200, 400}

5.3.2 Rossi

Para a modelagem dos dados dessa base foi usado somente o algoritmo **CoxPH-Fitter**, visto em 5.3.1.1, dado que a base é pequena e não faria sentido usar algoritmos de rede neural. O evento de interesse foi representado pela coluna **arrest** e o tempo do evento de interesse foi representado pela coluna **week**. Para evitar estratificação, não usamos a variável *age*.

Muito comum em modelos de regressão, o intercepto não é necessário ao usar Cox, dado que a função de risco base cumpre esse papel. Nesse algoritmo também é possível usar um parâmetro denominado *formula*, que permite passar uma equação linear específica para o modelo. Dada que a base é pequena fizemos iterações um a um, dois a dois e assim por diante. No entanto, as interações não fizeram diferença nos resultados e por isso optamos por usar a opção padrão do algoritmo na modelagem final. Importante ressaltar que várias variações são possíveis na fórmula no entanto não esgotamos as possibilidades neste trabalho dado que não é o escopo do mesmo. Os resultados serão discutidos na Seção 5.4.

5.3.3 COMPAS

No COMPAS também só foi utilizado o algoritmo **CoxPHFitter**, o mesmo usado na análise da ProPublica. O evento de interesse foi representado pela coluna **event** e o tempo do evento de interesse foi representado pela coluna **duration**. A partir do parâmetro *formula* foram testados 2 cenários com as seguintes variáveis: *score_text* e a fórmula $race + score_text + race * score_text$ que também foram usados na análise original. Os resultados serão discutidos na Seção 5.4.

5.4 Resultados

Passadas as etapas de modelagem e otimização, passamos para a avaliação das métricas de sucesso e escolha do melhor modelo. A seguir são apresentados os resultados

para cada uma das bases.

5.4.1 MIMIC-III

Seguimos com os seguintes passos ao trabalhar com modelos de aprendizado de máquina: 1. geramos cinquenta sementes diferentes; 2. achamos os conjuntos dos melhores parâmetros para cada um deles e 3. executamos os algoritmos com os parâmetros otimizados para cada semente. Após cada execução anotou-se o resultado do c-index dos dados de validação e teste.

Após ter todos os valores de c-index, compilou-se para cada algoritmo os valores mínimo, primeiro quartil, terceiro quartil e máximo. Ao final foi gerado um boxplot, presente na Figura 5.14, para comparar os resultados entre os algoritmos. Desse comparativo foi escolhido o modelo, que usamos para analisar os casos propostos no Capítulo 4.

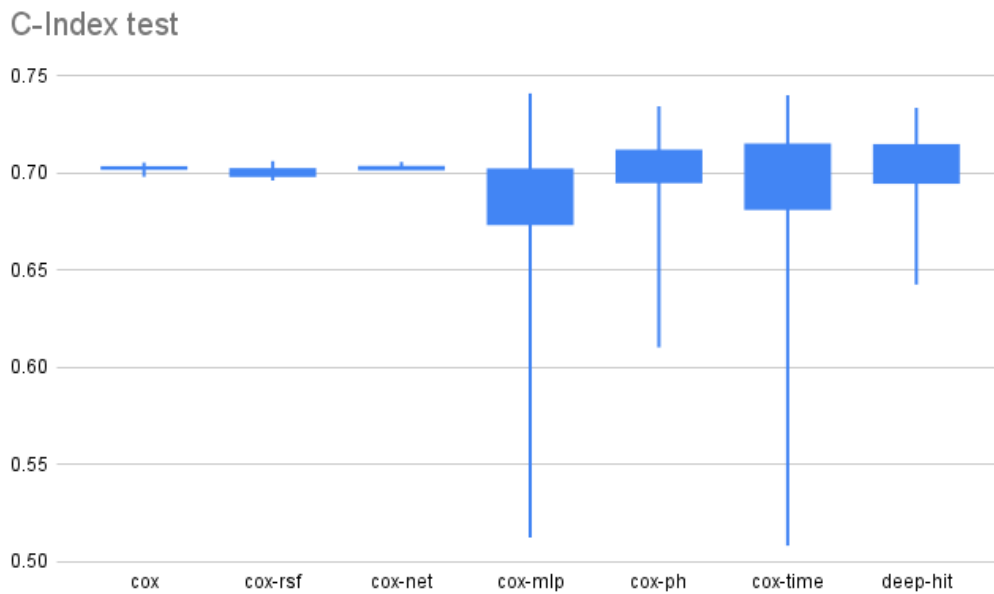


Figura 5.14. Boxplot comparativo entre os algoritmos

O algoritmo escolhido foi o Cox-Time, proposto por Kvamme et al. (2019). A escolha se deu por ser uma rede neural profunda, proposta recentemente, que permite contornar as restrições do risco proporcional. Além disso, teve bons resultados nos estudos propostos pelos autores e resultados satisfatórios nos nossos testes. Com isso o

próximo passo foi criar as amostras (grupos) e as funções com a aplicação das propostas vistas no Capítulo 4.

5.4.2 Rossi

Para a base Rossi optamos por apenas usar o algoritmo de Cox presente na biblioteca Lifelines, dado que a base possui poucos dados. O resultado da modelagem pode ser vista na Figura 5.15.

model		lifelines.CoxPHFitter									
duration col	'week'										
event col	'arrest'										
baseline estimation	breslow										
number of observations	432										
number of events observed	114										
partial log-likelihood	-662.74										
time fit was run	2021-07-29 11:55:46 UTC										
	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)	
fin	-0.42	0.66	0.19	-0.79	-0.05	0.45	0.95	-2.20	0.03	5.17	
race	0.28	1.32	0.31	-0.33	0.88	0.72	2.41	0.90	0.37	1.44	
wexp	-0.35	0.71	0.20	-0.75	0.05	0.47	1.05	-1.71	0.09	3.52	
mar	-0.55	0.58	0.38	-1.29	0.20	0.27	1.22	-1.45	0.15	2.75	
paro	-0.02	0.98	0.19	-0.40	0.36	0.67	1.43	-0.10	0.92	0.12	
prio	0.09	1.09	0.03	0.03	0.14	1.03	1.16	3.14	<0.005	9.19	
Concordance	0.63										
Partial AIC	1337.48										
log-likelihood ratio test	25.28 on 6 df										
-log2(p) of ll-ratio test	11.69										

Figura 5.15. Estatística sobre o resultado do modelo de Cox na base Rossi

Além do resultado do c-index, que ficou em 0.63 e é considerado um resultado de qualidade moderada, também é importante interpretar os coeficientes para entender o impacto que eles tem no risco da ocorrência do evento.

Vamos analisar as variáveis sensíveis *mar* (casado), *fin* (financiado) e *race* (raça). Considere primeiro a variável *mar*, ela possui um valor binário 0 ou 1, representando não casados e casados e um coeficiente de -0.55 . Esse valor de $\exp(-0.55) = 0.58$ está associado a taxa de risco dos presos que são casados, portanto significa uma diminuição de 42% no risco da ocorrência do evento.

O mesmo ocorre para a variável *fin*, que também possui valor binário 0 ou 1. O coeficiente possui valor -0.42 e $\exp(-0.42) = 0.66$, associado a taxa de risco com pre-

presos que receberam o financiamento. Isso representa uma redução de 36% no risco da ocorrência do evento. Por fim a variável *race* possui um coeficiente de 0.28 e $\exp(0.28) = 1.32$, ou seja, um aumento no risco de 32% quando associado com presos que são negros.

Na Figura 5.16 é possível ver o impacto de cada uma dessas variáveis considerando seu valor associado. No caso da variável *mar*, presos casados possuem um curva com menor risco associado, o mesmo acontece com a variável *fin*, onde presos que receberam o financiamento também possuem uma curva de sobrevivência com menor risco. No caso da variável *race*, presos negros possuem uma curva com maior risco associado.

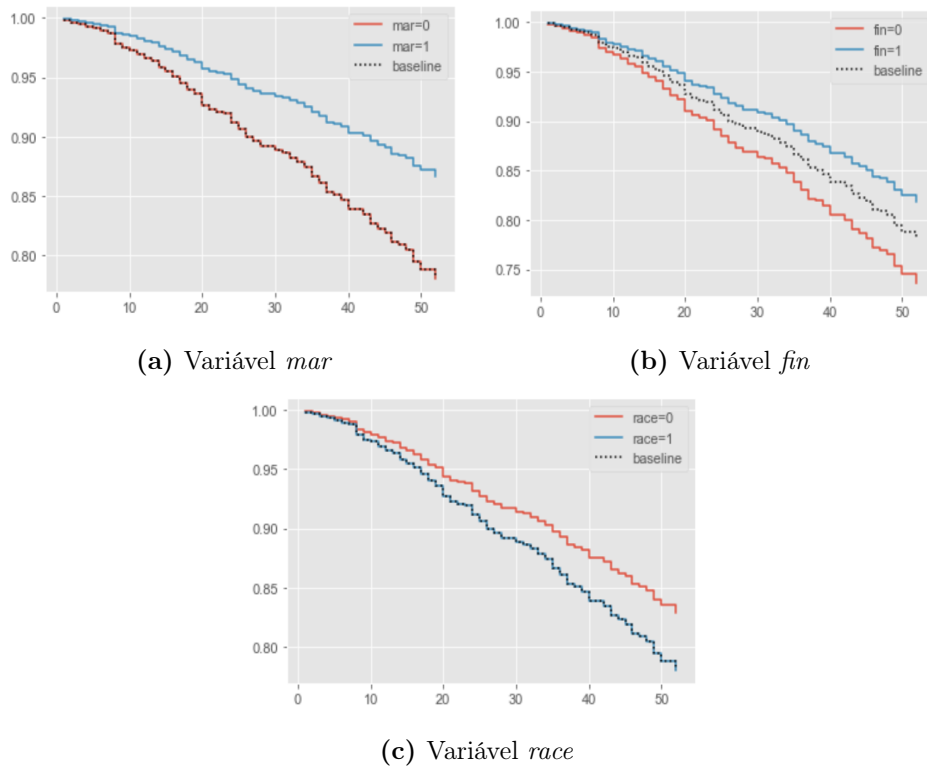


Figura 5.16. Impacto das variáveis na curva de sobrevivência

5.4.3 COMPAS

Para essa base também optamos por apenas usar o algoritmo de Cox presente na biblioteca Lifelines. Como foram testados dois cenários, os resultados da modelagem pode ser vistos nas Figuras 5.17 e 5.18

model		lifelines.CoxPHFitter									
duration col		'duration'									
event col		'event'									
baseline estimation		breslow									
number of observations		10314									
number of events observed		2759									
partial log-likelihood		-24097.25									
time fit was run		2021-09-21 02:25:03 UTC									
	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)	
score_text[T.factor_b_medium]	0.79	2.20	0.05	0.70	0.88	2.01	2.40	17.43	<0.005	223.73	
score_text[T.factor_c_high]	1.19	3.30	0.05	1.10	1.29	3.01	3.62	25.26	<0.005	465.13	

Concordance	0.63
Partial AIC	48198.51
log-likelihood ratio test	681.06 on 2 df
-log2(p) of ll-ratio test	491.28

Figura 5.17. Cenário 1 - Estatística sobre o resultado do modelo de Cox na base COMPAS

Para o cenário 1, o resultado do c-index ficou em 0.63, o mesmo obtido com a base Rossi. A variável *score_text*, usada na fórmula, está associada ao fator de risco dado pela ferramenta. Novamente vamos interpretar os coeficientes para entender o impacto que eles tem no risco da ocorrência do evento. Lembrando que temos três tipos de risco: baixo, médio e alto. A primeira variável é usada como base e as demais serão analisadas para entender se houve um aumento ou diminuição do risco da ocorrência do evento.

Considere o caso da variável de risco médio (*score_text/T.factor_b_medium*), ela possui um coeficiente de 0.79 e um valor $\exp(0.79) = 2.20$, ou seja, um aumento no risco de 120%. No caso da variável de risco alto (*score_text/T.factor_c_high*), o coeficiente é de 1.19 e um valor $\exp(1.19) = 3.30$, o que representa um aumento de 230% no risco!

Para o cenário 2, o resultado do c-index foi de 0.64, o que representa uma pequena melhora. A fórmula usada possui as variáveis de raça e fator de risco associadas. Em relação ao coeficientes usamos como base pessoas brancas com fator de risco baixo, que é o mesmo utilizado no estudo original.

Devido a quantidade de variáveis nesse modelo, vamos analisar somente 2 casos. Considerando *race[T.African-American]* temos um coeficiente de 0.29 e valor de

model		lifelines.CoxPHFitter										
duration col	'duration'											
event col	'event'											
baseline estimation	breslow											
number of observations	10314											
number of events observed	2759											
partial log-likelihood	-24072.72											
time fit was run	2021-09-21 02:25:05 UTC											
	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)		
race[T.African-American]	0.29	1.34	0.07	0.16	0.42	1.18	1.53	4.38	<0.005	16.38		
race[T.Asian]	-1.32	0.27	0.71	-2.71	0.06	0.07	1.07	-1.87	0.06	4.02		
race[T.Hispanic]	0.00	1.00	0.10	-0.20	0.20	0.82	1.23	0.04	0.97	0.04		
race[T.Native American]	-12.38	0.00	297.20	-594.87	570.12	0.00	3.97e+247	-0.04	0.97	0.05		
race[T.Other]	-0.05	0.95	0.12	-0.30	0.19	0.74	1.21	-0.44	0.66	0.60		
score_text[T.factor_b_medium]	0.83	2.29	0.08	0.67	0.99	1.96	2.68	10.37	<0.005	81.23		
score_text[T.factor_c_high]	1.28	3.59	0.10	1.09	1.47	2.97	4.33	13.30	<0.005	131.72		
race[T.African-American]:score_text[T.factor_b_medium]	-0.16	0.85	0.10	-0.36	0.04	0.70	1.04	-1.59	0.11	3.17		
race[T.Asian]:score_text[T.factor_b_medium]	1.28	3.58	0.92	-0.52	3.07	0.59	21.58	1.39	0.16	2.61		
race[T.Hispanic]:score_text[T.factor_b_medium]	-0.08	0.92	0.19	-0.44	0.29	0.64	1.33	-0.42	0.67	0.57		
race[T.Native American]:score_text[T.factor_b_medium]	12.39	2.41e+05	297.20	-570.11	594.89	0.00	2.27e+258	0.04	0.97	0.05		
race[T.Other]:score_text[T.factor_b_medium]	-0.17	0.84	0.25	-0.66	0.32	0.51	1.38	-0.69	0.49	1.02		
race[T.African-American]:score_text[T.factor_c_high]	-0.26	0.77	0.11	-0.49	-0.04	0.62	0.96	-2.30	0.02	5.54		
race[T.Asian]:score_text[T.factor_c_high]	2.16	8.68	0.92	0.36	3.96	1.44	52.45	2.35	0.02	5.75		
race[T.Hispanic]:score_text[T.factor_c_high]	-0.15	0.86	0.22	-0.58	0.27	0.56	1.32	-0.70	0.48	1.06		
race[T.Native American]:score_text[T.factor_c_high]	12.54	2.80e+05	297.20	-569.96	595.04	0.00	2.64e+258	0.04	0.97	0.05		
race[T.Other]:score_text[T.factor_c_high]	0.46	1.59	0.28	-0.08	1.01	0.92	2.74	1.66	0.10	3.35		
Concordance	0.64											
Partial AIC	48179.44											
log-likelihood ratio test	730.12 on 17 df											
-log2(p) of ll-ratio test	476.58											

Figura 5.18. Cenário 2 - Estatística sobre o resultado do modelo de Cox na base COMPAS

$\exp(0.29) = 1.34$, ou seja, um aumento de 34% no risco de ocorrência do evento. No entanto se analisarmos $race[T.African-American]:score_text[T.factor_c_high]$ o coeficiente possui valor negativo de -0.26 e $\exp(-0.26) = 0.77$. Isso representa uma redução de 23% no risco da ocorrência do evento.

Todo o tratamento dos dados, implementações dos algoritmos, conjunto de melhores parâmetros testados e arquivos com resultados que foram comentados ao longo deste capítulo podem ser encontrados no repositório desse projeto.

Neste capítulo apresentamos as bases de dados utilizadas (MIMIC-III, Rossi e COMPAS), a caracterização dos dados junto com as análises exploratórias e os detalhes envolvendo as modelagens de cada problema. Após o treinamento dos modelos, grupos de teste considerando a variável sensível serviram para avaliar a performance e a justiça dos modelos. No Capítulo 6 iremos apresentar os gráficos e resultados dessa avaliação em conjunto com as discussões de justiça para cada uma das quatro propostas sugeridas nesta dissertação.

Capítulo 6

Discussões sobre justiça

No Capítulo 4 foram apresentadas três propostas de adaptação das definições de justiça para modelos de análise de sobrevivência e uma nova métrica denominada justiça de filas para verificar a existência de injustiça em casos de comparação de grupos. No Capítulo 5 apresentamos as bases de dados e o detalhamento de todos os passos até a modelagem. Por fim, com o modelo treinado, usamos os grupos de teste para a avaliação das métricas. Neste capítulo iremos mostrar os resultados e entrar nas discussões de justiça para cada uma das propostas sugeridas nesta dissertação.

6.1 Proposta 1: Divergência em paridade demográfica

Conforme a proposta apresentada na subseção 4.1.1, buscamos curvas de sobrevivência para ambos os valores da variável sensível, tal que as curvas sejam do tipo: $(S(t|z = 0) = S(t|z = 1)) \approx (\hat{S}(t|z = 0) = \hat{S}(t|z = 1))$.

Para fazer as comparações das curvas, primeiro usamos o método de Kaplan-Meier (KM) apresentado na Seção 2.2.1.1. A ideia é obter a função de sobrevivência de cada grupo a partir dos dados coletados. Além disso, aplicamos o teste estatístico Logrank que compara as probabilidades de sobrevivência de cada curva entre os diferentes grupos. Neste teste, a hipótese nula é que não há diferença entre os grupos e a hipótese alternativa é que há diferença entre os grupos.

6.1.1 MIMIC-III

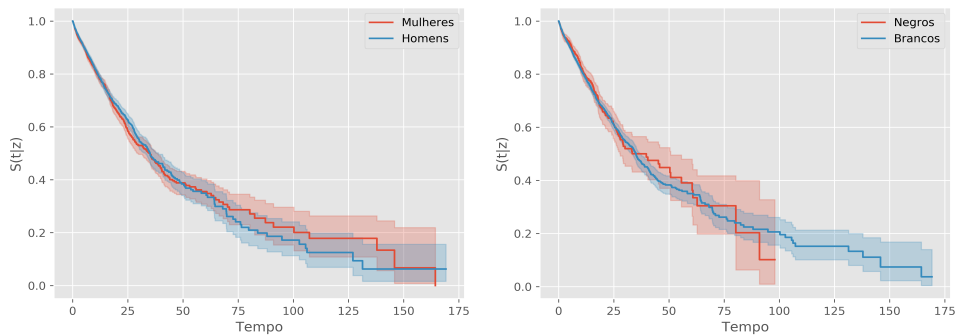
O modelo com o algoritmo Cox-Time foi treinado com 3851 amostras, validado com 963 amostras e o percentual de censura foi de aproximadamente 30% das amostras.

Das 2901 amostras de teste disponíveis, foram criados grupos levando em consideração duas variáveis sensíveis: gênero e raça. Para gênero usamos apenas mulher e homem e em relação a raça usamos negros e brancos. O percentual de censura na amostra de teste foi de aproximadamente 16%.

Nesse primeiro caso foram feitas quatro análises:

- mulheres / homens
- negros / brancos
- mulheres negras / mulheres brancas
- homens negros / homens brancos

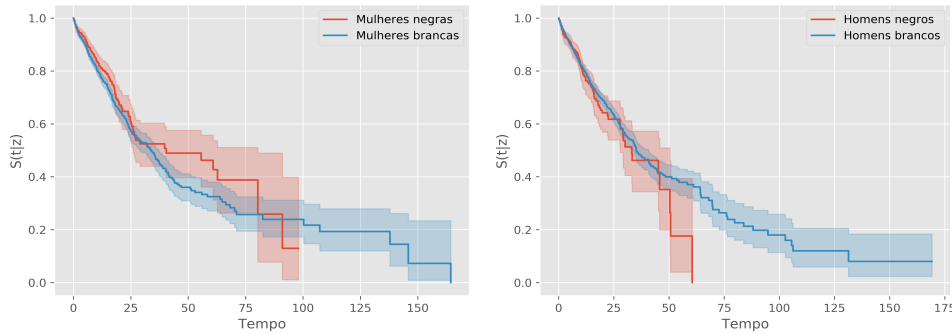
Nas Figuras 6.1 e 6.2 plotamos as curvas considerando apenas as variáveis de gênero e raça. Analisando as curvas KM da Figura 6.1 não há diferença significativa entre elas enquanto na Figura 6.2 parece existir uma diferença. Exclusivamente pelas imagens fica difícil verificar e confirmar se há diferença significativa entre as curvas, por isso o teste Logrank é essencial. Importante salientar que há uma limitação na diversidade de amostras e por não possuir dados observados para toda a janela de tempo as curvas para pessoas negras acabam abruptamente.



(a) Curva Kaplan-Meier para gênero

(b) Curva Kaplan-Meier para raça

Figura 6.1. Curvas de sobrevivência divididas em gênero e raça



(a) Curva Kaplan-Meier para mulheres negras e brancas (b) Curva Kaplan-Meier para homens negros e brancos

Figura 6.2. Curvas de sobrevivência para gênero/raça

Na Tabela 6.1 os resultados do teste estatístico mostram que apenas a curva 6.2 (a) possui um valor-p menor que 10% chegando perto do limiar de 5%, ou seja, rejeitando a hipótese nula. Desta forma, pode-se concluir que há diferença significativa entre as curvas KM para mulheres negras e brancas.

Tabela 6.1. Tabela com resultados do teste LogRank para a proposta 1 da base MIMIC-III

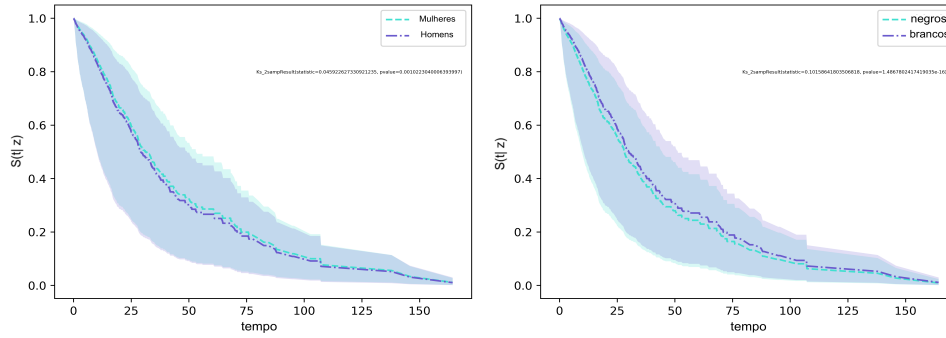
Grupos	LR	Valor-p
Mulheres / Homens	0.575	0.448
Negros / Brancos	0.650	0.419
Mulheres negras / mulheres brancas	3.609	0.057
Homens negros / homens brancos	0.619	0.431

Em seguida, com o uso do modelo de Cox treinado, fizemos a predição da curva de sobrevivência para cada um desses mesmos grupos. Para todas as bases, separamos os resultados em **justos** e **injustos** para facilitar o entendimento. De maneira geral, nem todos os resultados serão apresentados neste capítulo devido a grande quantidade de gráficos existentes. O que não for discutido aqui estará disponível no Apêndice A.

Para a análise dos resultados das curvas preditas aplicou-se o teste estatístico de Kolmogorov-Smirnov (KS) (Hodges, 1958), que compara duas distribuições contínuas e independentes $F(x)$ e $G(x)$ para todos os valores de x , onde $x \in \mathbb{R}$. Para esse teste a hipótese nula é que as distribuições são idênticas e a hipótese alternativa é que elas não são idênticas.

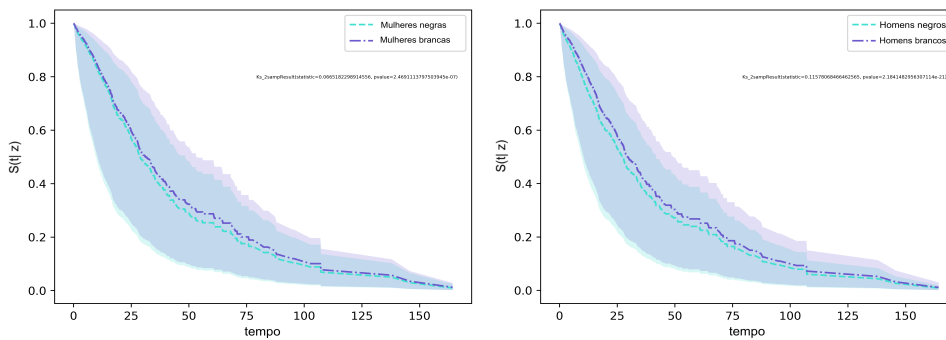
Os casos considerados justos foram aqueles nos quais não houve disparidade entre curvas de sobrevivência observadas nos dados quando comparadas com previsões. E

os casos injustos foram curvas que apresentaram comportamentos distintos. A Tabela 6.3 apresenta o resultado para cada caso estudado.



(a) Curva predita para mulheres e homens (b) Curva predita para negros e brancos

Figura 6.3. Curva predita para gênero e raça



(a) Curva predita para mulheres negras e brancas (b) Curva predita para homens negros e brancos

Figura 6.4. Curva predita para gênero-raça

Os gráficos das Figuras 6.3 e 6.4 mostram as curvas de sobrevivência de cada grupo e a faixa do intervalo de confiança de 95%. A tabela 6.2 traz os resultados do teste KS.

Tabela 6.2. Resultados do teste KS para a proposta 1 da base MIMIC-III

Grupos	KS	Valor-p
Mulheres / Homens	0.05	0.001
Negros / Brancos	0.10	$1.48e^{-16}$
Mulheres negras / mulheres brancas	0.07	$2.47e^{-7}$
Homens negros / homens brancos	0.12	$2.18e^{-21}$

Analisando os gráficos das curvas preditas percebe-se que todas as curvas possuem comportamentos análogos. Em relação ao teste KS, o valor-p foi menor que 5% em todos os casos, ou seja, a hipótese nula foi rejeita e portanto as distribuições não são idênticas. Apesar de a quantidade de pessoas variar bastante, a proporção de casos de censura em cada grupo foi a mesma, $\simeq 16\%$.

Ao avaliar os resultados, houve um caso de disparidade entre as curvas KM e predita, que foi para mulheres negras e brancas, onde a curva empírica apresentou uma diferença que não foi capturada na predição. Logo, esse caso não satisfaz a definição de divergência em paridade demográfica proposta. Não fica evidente, analisando o gráfico da curva KM, se essa diferença representa um viés para mulheres negras ou brancas, dado que as curvas se cruzam ao longo do tempo. No entanto, é nítido que as curvas são diferentes. Aqui, a limitação de dados ao longo do tempo acaba prejudicando uma análise mais profunda dos resultados. Nos demais casos não houve diferença. Os resultados resumidos encontram-se na Tabela 6.3 .

Tabela 6.3. Resultados para a proposta de divergência em paridade demográfica

Grupos	Resultado
Mulheres / Homens	Justo
Negros / Brancos	Justo
Mulheres negras / mulheres brancas	Injusto
Homens negros / homens brancos	Justo

6.1.2 Rossi

Para a proposta 1 obtivemos as curvas empíricas e preditas com foco para três tipos de análises: raça, casados e financiados. Retomando as estatísticas sobre cada grupo: são 379 negros e 53 de outras raças; 379 não casados e 53 casados; 216 financiados e 216 não financiados e sob os resultados foram aplicados os testes estatísticos LogRank e Kolmogorov-Smirnov.

Os casos considerados justos foram aqueles nos quais não houve disparidade entre curvas de sobrevivência observadas nos dados quando comparadas com previsões. E os casos injustos foram curvas que apresentaram comportamentos distintos. Os resultados para essa proposta estão na tabela 6.4.

As curvas de sobrevivência usando o estimador de Kaplan-Meier para cada variável são mostradas na Figura 6.5. E os resultados do teste estatístico LogRank estão na Tabela 6.5.

Tabela 6.4. Resultados para a proposta de divergência em paridade demográfica na base Rossi

Grupos	Resultado
Casados / não casados	Justo
Financiados / não financiados	Justo
Negros / brancos	Justo

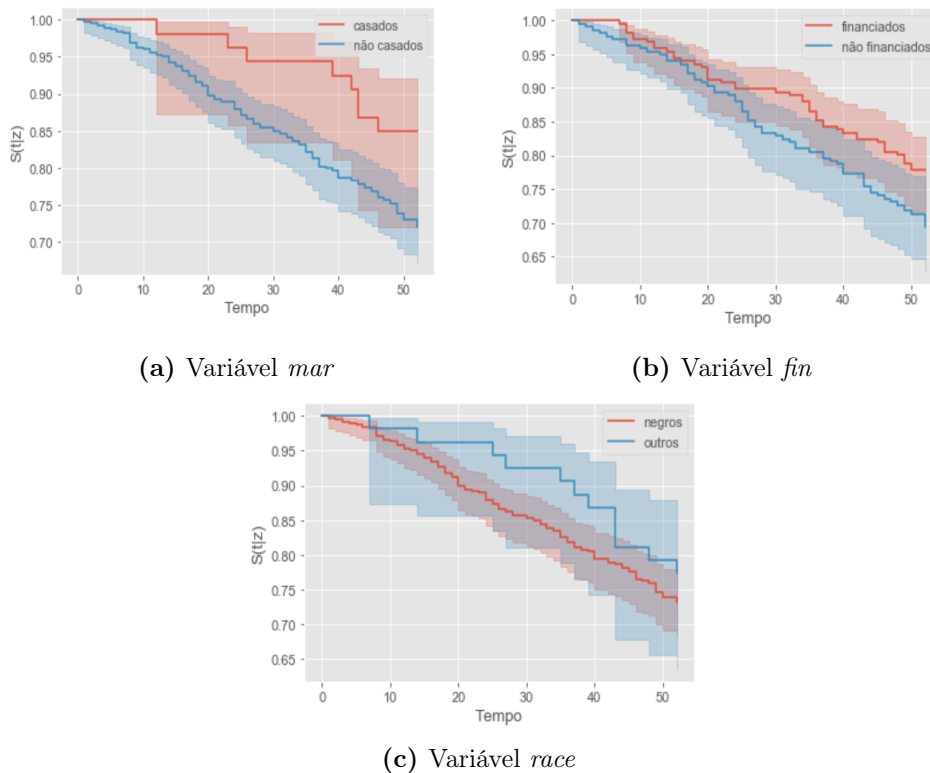


Figura 6.5. Curva de sobrevivência com o estimador Kaplan-Meier

Tabela 6.5. Tabela com resultados do teste LogRank para a base Rossi

Grupos	LR	Valor-p
Casados / não casados	3.937	0.047
Financiados / não financiados	3.837	0.050
Negros / Outros	0.576	0.447

Apesar das curvas empíricas mostrarem diferenças relevantes em cada caso, os resultados do teste Logrank mostram que o valor-p foi significativo apenas para os

grupos *mar* e *fin* e não significativo no caso da variável *race*.

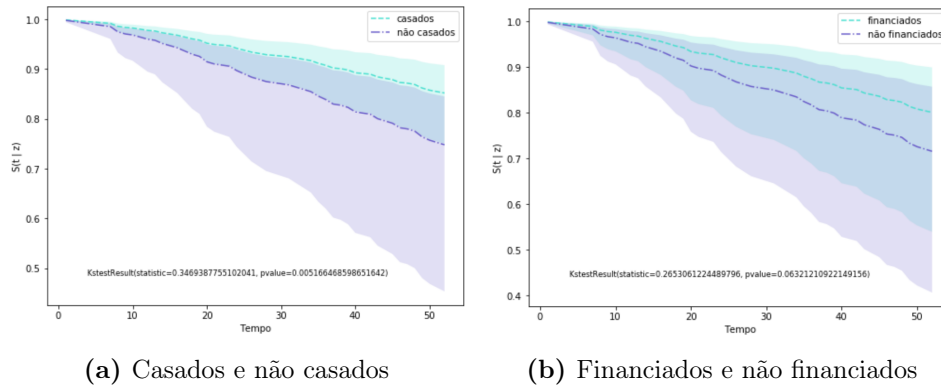


Figura 6.6. Paridade demográfica

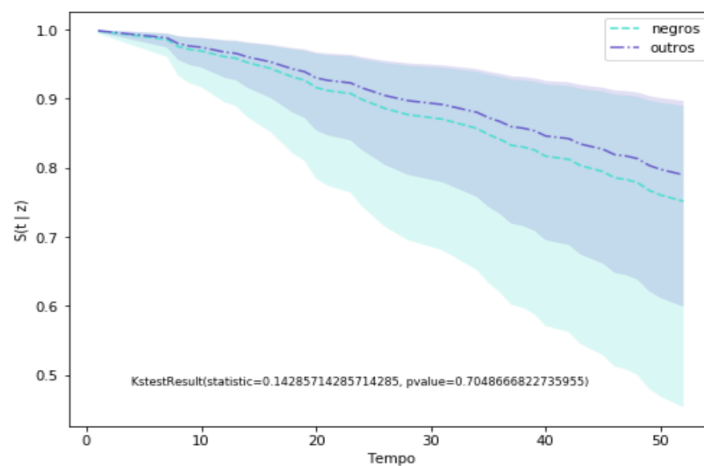


Figura 6.7. Paridade demográfica considerando raça

No caso das curvas preditas, os gráficos das Figuras 6.6 e 6.7 mostram a curva de sobrevivência para cada grupo e a faixa do intervalo de confiança de 95%. A tabela 6.6 traz os resultados do teste KS.

Analisando o gráfico da Figura 6.7, as curvas possuem comportamento semelhante. Já as curvas da Figura 6.6 apresentam diferença, tal que elas começam a divergir após a décima semana e até o final do estudo os presos de cada grupo apresentam probabilidades de sobrevivência distintas. Em relação ao teste KS, o valor-p foi menor que 5% nos casos injustos e maior que 5% no caso de gênero. Neste caso o teste indica que as distribuições são iguais mas como o caso foi considerado justo, a definição proposta foi satisfeita.

Tabela 6.6. Resultados do teste KS para proposta 2 da base Rossi

Grupos	KS	Valor-p
Casados / Não casados	0.35	0.005
Financiados / Não financiados	0.26	0.06
Negros / Outros	0.14	0.70

Ao avaliar os resultados apresentados nesta subseção, os casos não apresentaram disparidade entre as curvas KM e preditas, ou seja, a curva empírica e a curva predita estão de acordo. Logo, esse casos satisfazem a proposta de divergência em paridade demográfica. Os resultados resumidos encontram-se na Tabela 6.12.

6.1.3 COMPAS

Para essa base, tanto a curva de sobrevivência empírica quanto a predita tiveram seu foco na variável sensível raça. Foi considerado justo se as curvas de sobrevivência empíricas e preditas, quando comparadas, apresentaram comportamentos similares e injusto se as curvas que apresentaram comportamentos distintos.

Tabela 6.7. Resultados para a proposta de divergência em paridade demográfica na base COMPAS

Grupos	Resultado
Negros / brancos	Justo

As curvas de sobrevivência empíricas são mostrada na Figura 6.8. Analisando a figura fica nítido que pessoas negras possuem uma curva de sobrevivência que indica maior probabilidade de reincidência do que brancos. O resultado do teste LogRank, presente na Tabela 6.8 mostra que o valor-p é, de fato, significativo.

Tabela 6.8. Tabela com resultados do teste LogRank para a proposta 1 da base COMPAS

Grupos	LR	Valor-p
Negros / Brancos	95.28	$1.65e^{-22}$

Tabela 6.9. Tabela com resultados do teste KS para a proposta 1 da base COMPAS

Grupos	KS	Valor-p
Negros / Brancos	0.5	0 ¹

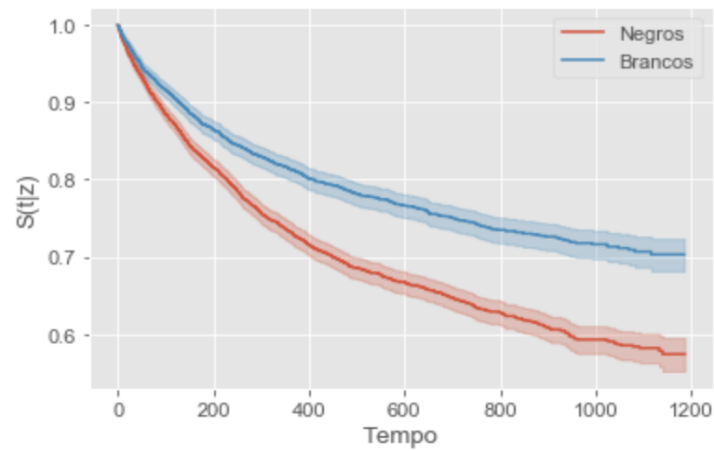


Figura 6.8. Curva de sobrevivência com o estimador Kaplan-Meier para a base COMPAS

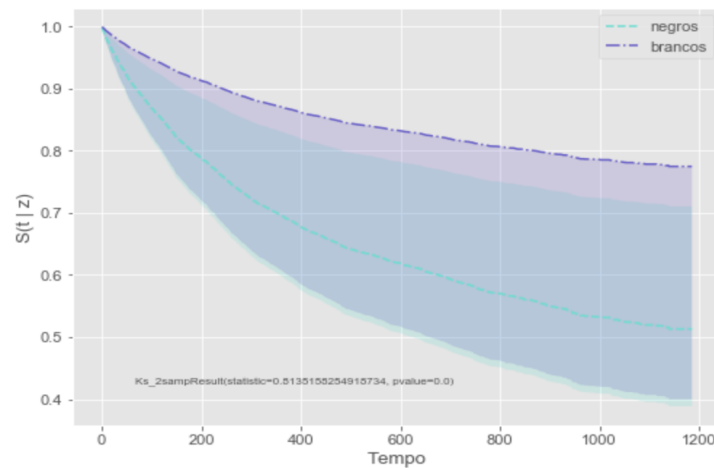


Figura 6.9. Curva predita considerando raça para a base COMPAS

O gráfico da Figura 6.9 mostra a curva de sobrevivência predita para cada grupo e a faixa do intervalo de confiança de 95%. A tabela 6.9 também traz os resultados do teste KS. Analisando o gráfico, as curvas possuem comportamentos distintos, dado que elas começam a divergir já no início da janela de tempo e no final da observação cada grupo apresenta probabilidades de sobrevivência bem diferentes. Em relação ao teste KS, o valor-p foi muito menor que 5% indicando que as distribuições não são iguais.

Ao avaliar os resultados apresentados nesta subseção, os casos não apresentaram disparidade entre as curvas KM e preditas, ou seja, a curva empírica e a curva predita estão de acordo. Logo, esses casos satisfazem a proposta de divergência em paridade

¹Valor extremamente pequeno

demográfica. Os resultados resumidos encontram-se na Tabela 6.7.

6.2 Proposta 2: Divergência em paridade demográfica condicionada

Conforme a proposta vista em 4.1.2, buscamos curvas de sobrevivência para ambos os valores da variável sensível, tal que a predição seja do tipo: $S(t|\mathbf{w}, z = 0) = S(t|\mathbf{w}, z = 1) \approx \hat{S}(t|\mathbf{w}, z = 0) = \hat{S}(t|\mathbf{w}, z = 1)$, onde temos atributos condicionados.

6.2.1 MIMIC-III

Para essa proposta criamos quarenta e oito grupos de teste dividindo entre gênero, raça e gênero-raça. Os fatores condicionais escolhidos foram a pontuação Oasis e quatro tipos de diagnóstico dados pelo hospital: câncer, diabetes, coração e transplante. O Alzheimer não foi escolhido por ter poucas ocorrências e a taxa de censura divergir dos demais casos. Cada um desses fatores foi condicionado individualmente.

As análises foram divididas em:

- Mulheres | Fator X / Homens | Fator X
- Negros | Fator X / Brancos | Fator X
- Mulheres negras | Fator X / Mulheres brancas | Fator X
- Homens negros | Fator X / Homens brancos | Fator X

Sendo o fator X: Oasis, câncer, diabetes, coração ou transplante.

Os casos considerados justos foram aqueles nos quais não houve disparidade entre curvas de sobrevivência observadas nos dados quando comparadas com previsões. E os casos injustos foram curvas que apresentaram comportamentos distintos. A Tabela 6.12 apresenta o resultado para cada caso estudado.

Os gráficos com as curvas Kaplan-Meier, que serão discutidos nesse capítulo, estão presentes na Figura 6.10 e os resultados do teste Logrank encontram-se na Tabela 6.10. Os demais gráficos e resultados, que não serão discutidos neste capítulo, podem ser encontrados no Apêndice A.

Analisando os gráficos da Figura 6.10 parece existir diferenças entre as curvas KM. No caso do fator condicional ser câncer, pessoas negras possuem uma curva empírica de sobrevivência pior do que pessoas brancas. Já nos casos de o fator ser a

pontuação Oasis ou transplante, pessoas brancas possuem uma curva empírica pior. Porém, é importante lembrar que há uma limitação na diversidade de amostras e por não possuir dados observados para toda a janela de tempo as curvas para pessoas negras acabam abruptamente.

Na Tabela 6.10 os resultados do teste estatístico LogRank mostram que todas as curvas possuem um valor-p menor que 5%, ou seja, rejeitando a hipótese nula. Desta forma, pode-se concluir que há diferença significativa entre as curvas KM para todos esses grupos.

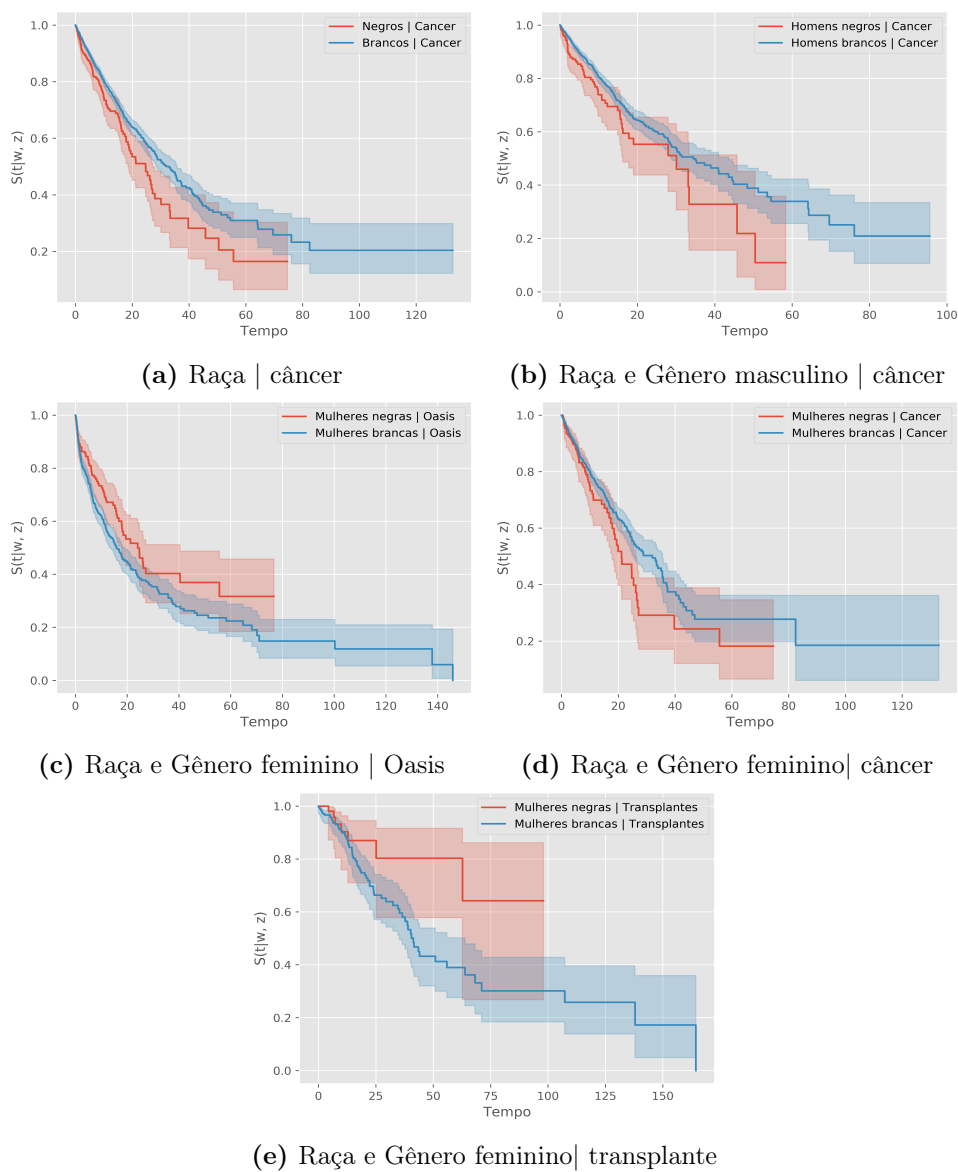


Figura 6.10. Curvas KM para a proposta 2 da base MIMIC-III

Tabela 6.10. Tabela com resultados do teste LogRank para a proposta 2 da base MIMIC-III

Grupos	LR	Valor-p
Negros / Brancos câncer	9.8	0.00017
Mulheres negras / mulheres brancas Oasis	6.24	0.012
Mulheres negras / mulheres brancas câncer	3.81	0.05
Mulheres negras / mulheres brancas transplante	3.8	0.051
Homens negros / homens brancos câncer	5.83	0.015

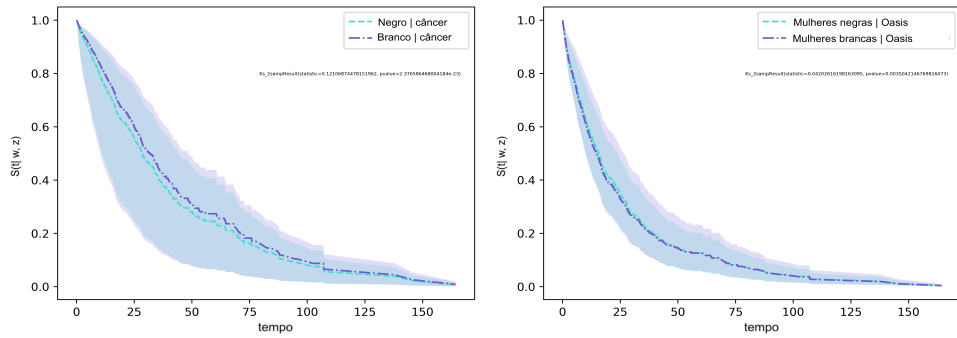
Tabela 6.11. Resultados do teste KS para a proposta 2 da base MIMIC-III

Grupos	KS	Valor-p
Negros câncer / Brancos câncer	0.12	$2.37e^{-23}$
Mulheres negras Oasis / mulheres brancas Oasis	0.04	0.003
Mulheres negras câncer / mulheres brancas câncer	0.15	$1.69e^{-34}$
Mulheres negras transplante / mulheres brancas transplante	0.10	$2.32e^{-17}$
Homens negros câncer / homens brancos câncer	0.14	$1.69e^{-34}$

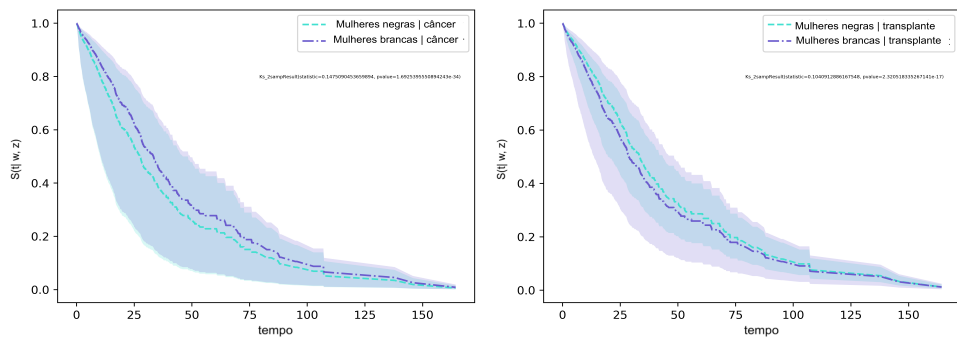
Seguindo para o modelo de Cox, os gráficos da Figura 6.11 mostram as curvas de sobrevivência preditas para cada grupo e a faixa do intervalo de confiança é de 95%. A tabela 6.11 traz os resultados do teste KS. Novamente, apenas alguns gráficos são apresentados nesta subseção, os demais estão disponíveis no Apêndice A.

Analisando os gráficos, todas as curvas preditas possuem comportamentos análogos. Em relação ao teste KS, o valor-p foi menor que 5% nos casos apresentados nos gráficos, ou seja, a hipótese nula foi rejeita e conseqüentemente as distribuições não são idênticas.

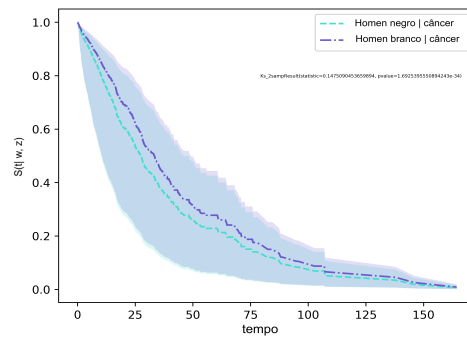
Ao avaliar os resultados apresentados nesta subseção, todos os casos apresentaram disparidade entre as curvas KM e preditas, ou seja, a curva empírica apresentou uma diferença que não foi capturada na predição. Logo, esse casos não satisfazem a definição de divergência em paridade demográfica proposta. Nos demais casos não houve diferença. O resultado para cada caso encontra-se na Tabela 6.12.



(a) Curva predita para negros e brancos dado câncer (b) Curva predita para mulheres negras e brancas dada a pontuação Oasis



(c) Curva predita para mulheres negras e brancas dado câncer (d) Curva predita para mulheres negras e brancas dado transplante



(e) Curva predita para homens negros e brancos dado câncer

Figura 6.11. Curva predita para a proposta 2 da base MIMIC-III

Tabela 6.12. Resultados para a proposta de divergência de paridade demográfica condicionada na base MIMIC-III

Grupos	Resultado
Mulheres Oasis / Homens Oasis	Justo
Mulheres câncer / Homens câncer	Justo
Mulheres coração / Homens coração	Justo
Mulheres diabetes / Homens diabetes	Justo
Mulheres transplante / Homens transplante	Justo
Negros Oasis / Brancos Oasis	Justo
Negros câncer / Brancos câncer	Injusto
Negros coração / Brancos coração	Justo
Negros diabetes / Brancos diabetes	Justo
Negros transplante / Brancos transplante	Justo
Mulheres negras Oasis / mulheres brancas Oasis	Injusto
Mulheres negras câncer / mulheres brancas câncer	Injusto
Mulheres negras coração / mulheres brancas coração	Justo
Mulheres negras diabetes / mulheres brancas diabetes	Injusto
Mulheres negras transplante / mulheres brancas transplante	Injusto
Homens negros Oasis / homens brancos Oasis	Justo
Homens negros câncer / homens brancos câncer	Injusto
Homens negros coração / homens brancos coração	Justo
Homens negros diabetes / homens brancos diabetes	Justo
Homens negros transplante / homens brancos transplante	Justo

6.2.2 Rossi

Para essa proposta criamos oito grupos de teste divididos entre raça e os fatores condicionais foram estado civil e financiamento:

- Negros | casados e financiados / Outros | casados e financiados
- Negros | casados e não financiados / Outros | casados e não financiados
- Negros | não casados e financiados / Outros | não casados e financiados
- Negros | não casados e não financiados / Outros | não casados e não financiados

Para esta proposta os casos considerados justos foram aqueles nos quais as curvas empíricas e preditas, quando comparadas, apresentaram comportamentos similares, ou seja, são praticamente iguais. E os casos injustos foram curvas que apresentaram comportamentos distintos. A curva KM para cada um desses grupos é mostrada na Figura 6.12.

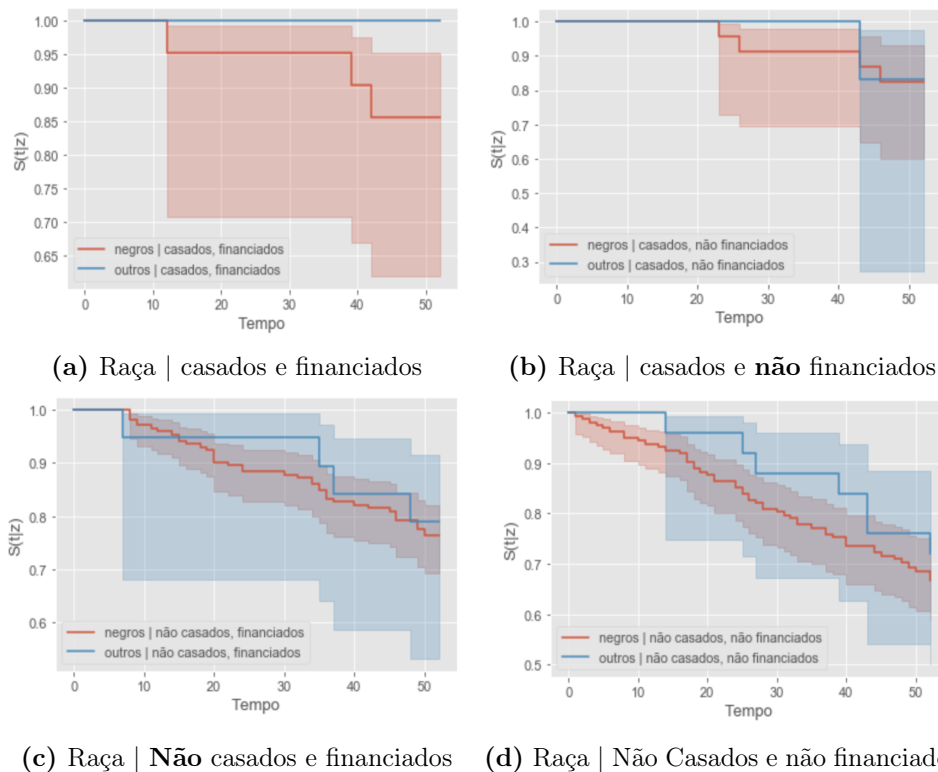


Figura 6.12. Curvas KM para a definição de divergência em paridade demográfica condicionada na base Rossi

Na Figura 6.12 (a) é nítida a diferença entre os dois grupos. Negros possuem probabilidade de sobrevivência menor ao longo do tempo enquanto o outro grupo mantém-se constante durante todo o período observado. Importante salientar que a amostra de pessoas casadas é muito menor que a de pessoas não casadas. Nas imagens (c) e (d) as curvas possuem mais pontos observados, sendo a de negros maior que a do outro grupo. Em ambas, a curva de sobrevivência média para pessoas negras é ligeiramente pior do que para o outro grupo.

No entanto, pelos resultados do teste LogRank, presentes na Tabela 6.13 nenhum deles possui um valor-p menor que 10% chegando perto do limiar de 5%, ou seja, rejeitando a hipótese nula. Desta forma, pode-se concluir que não há diferença significativa entre as curvas.

Tabela 6.13. Tabela com resultados do teste LogRank para a proposta 2 da base Rossi

Grupos	LR	Valor-p
Negros casados e financiados / Outros casados e financiados	0.45	0.5
Negros casados e não financiados / Outros casados e não financiados	0.005	0.94
Negros não casados e financiados / Outros não casados e financiados	0.07	0.78
Negros não casados e não financiados / Outros não casados e não financiados	0.41	0.52

Seguindo a análise dos resultados, os gráficos da Figura 6.13 mostram a curva de sobrevivência predita de cada grupo e a faixa do intervalo de confiança de 95%. A tabela 6.14 traz os resultados do teste KS. Analisando os gráficos, as curvas possuem comportamentos similares. Em relação ao teste KS, o valor-p foi maior que 5% nos casos apresentados, não sendo possível rejeitar a hipótese nula.

Ao avaliar os resultados apresentados nesta subseção, os casos não apresentaram disparidade entre as curvas KM e preditas, ou seja, a curva empírica e a curva predita estão de acordo. Logo, esse casos satisfazem a proposta de divergência em paridade demográfica condicionada. Os resultados resumidos encontram-se na Tabela 6.15.

Apesar de a proposta de divergência em paridade demográfica condicionada não ter casos considerados injustos, dado os critérios propostos, a análise exploratória mostrou que raça é uma variável que influencia o risco de ocorrência do evento. Dessa maneira, é importante o estudo de outras definições de justiça aplicadas a análise de sobrevivência que possam captar esse impacto.

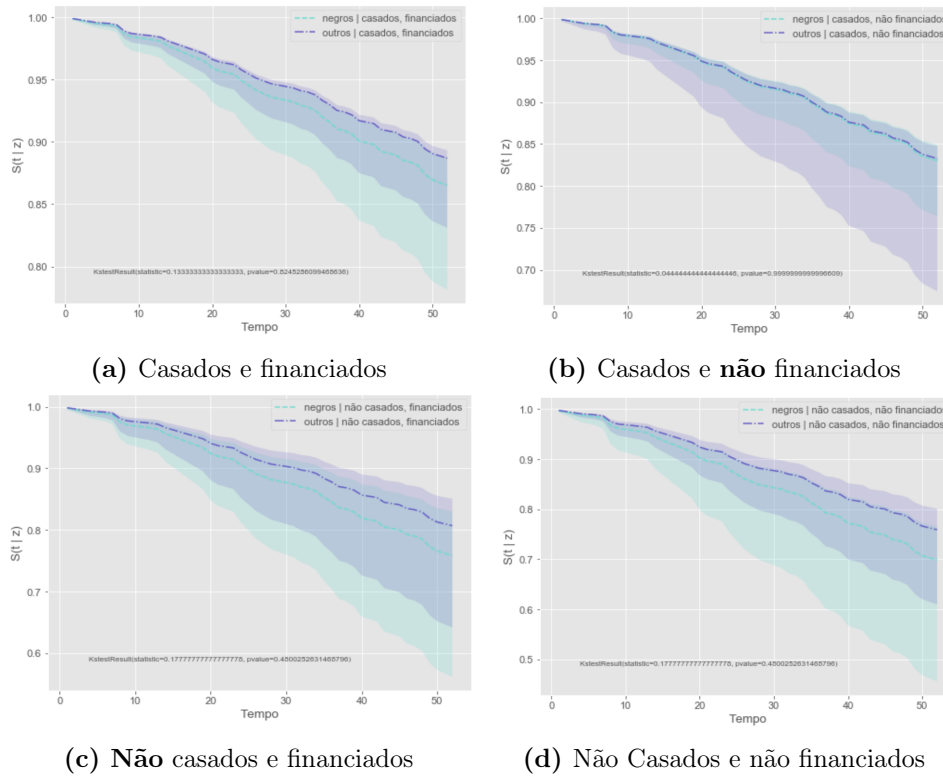


Figura 6.13. Paridade demográfica condicionada na base Rossi

Tabela 6.14. Resultados do teste KS para proposta de divergência em paridade demográfica condicionada na base Rossi

Grupos	KS	Valor-p
Negros casados e financiados / Outros casados e financiados	0.13	0.52
Negros casados e não financiados / Outros casados e não financiados	0.04	0.99
Negros não casados e financiados / Outros não casados e financiados	0.18	0.45
Negros não casados e não financiados / Outros não casados e não financiados	0.18	0.45

Tabela 6.15. Resultados para a proposta de divergência em paridade demográfica condicionada na base Rossi

Grupos	Resultado
Negros casados e financiados / Outros casados e financiados	Justo
Negros casados e não financiados / Outros casados e não financiados	Justo
Negros não casados e financiados / Outros não casados e financiados	Justo
Negros não casados e não financiados / Outros não casados e não financiados	Justo

6.2.3 COMPAS

Nesta proposta buscamos curvas de sobrevivência para ambos os valores da variável sensível, tal que a predição seja do tipo: $S(t|\mathbf{w}, z = 0) = S(t|\mathbf{w}, z = 1) \approx \hat{S}(t|\mathbf{w}, z = 0) = \hat{S}(t|\mathbf{w}, z = 1)$, onde temos atributos condicionados. Para esse caso criamos três grupos de teste usando a variável sensível raça e o atributo condicional escolhido foi fator de risco:

- Negros | risco baixo / Brancos | risco baixo
- Negros | risco médio / Brancos | risco médio
- Negros | risco alto / Brancos | risco alto

Os casos considerados justos foram aqueles nos quais as curvas de sobrevivência empírica e predita, quando comparadas, apresentaram comportamentos similares. E os casos injustos foram curvas que apresentaram comportamentos distintos.

As curvas de sobrevivência empíricas são mostrada na Figura 6.14. Analisando o gráfico (a) fica nítido que negros com risco baixo possuem maior probabilidade de reincidência do que brancos com risco baixo. Enquanto nas curvas para risco médio e alto essa diferença parece não existir. O resultado do teste LogRank, presente na Tabela 6.16 mostra que o valor-p para o gráfico (a) é, de fato, significativo.

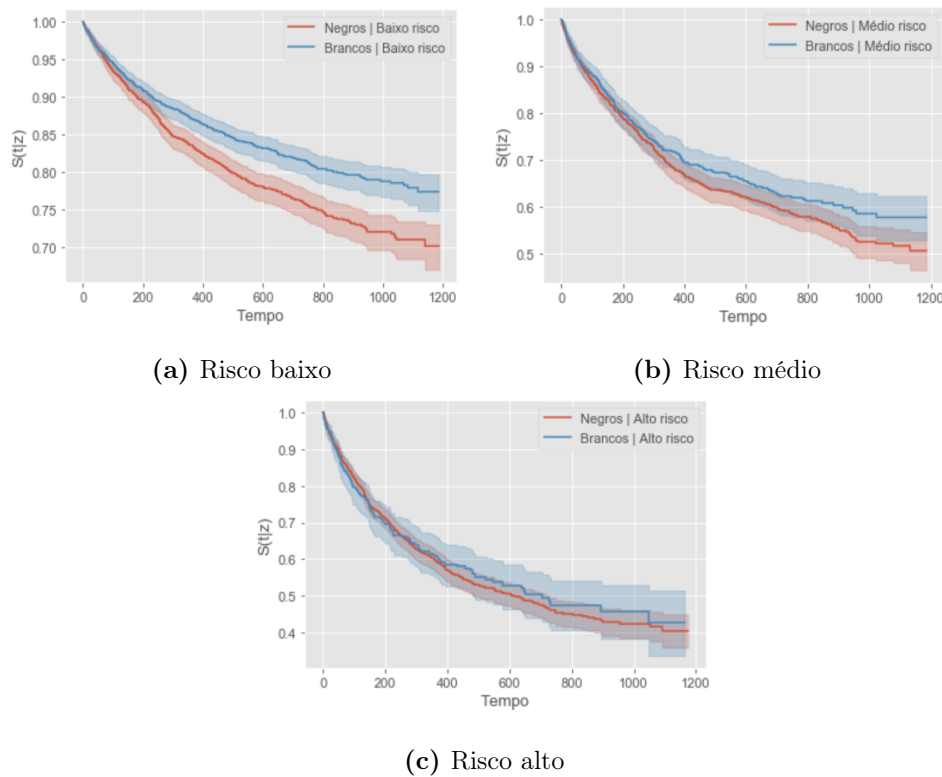


Figura 6.14. Curva KM para proposta 2 na base COMPAS

Tabela 6.16. Resultados do teste LogRank para proposta 2 na base COMPAS

Grupos	LR	Valor-p
Negros risco baixo / Brancos risco baixo	19.5	$1.00e^{-5}$
Negros risco médio / Brancos risco médio	2.94	0.08
Negros risco alto / Brancos risco alto	0.10	0.74

Os gráficos da Figura 6.15 mostram a curva de sobrevivência predita de cada grupo e a faixa do intervalo de confiança de 95%. A tabela 6.17 traz os resultados do teste KS. Analisando o gráfico de risco baixo, as curvas divergem ao longo do tempo, fazendo com que as probabilidades de sobrevivência sejam diferentes ao final da janela de tempo observada. Para as curvas de risco médio e alto os comportamentos foram considerados similares. Em relação ao teste KS, o valor-p foi menor que 5% nos casos apresentados, logo, a hipótese nula foi rejeita e portanto as distribuições não são idênticas.

Ao avaliar os resultados apresentados nesta subseção, os casos não apresentaram disparidade entre as curvas KM e preditas, ou seja, a curva empírica e a curva predita

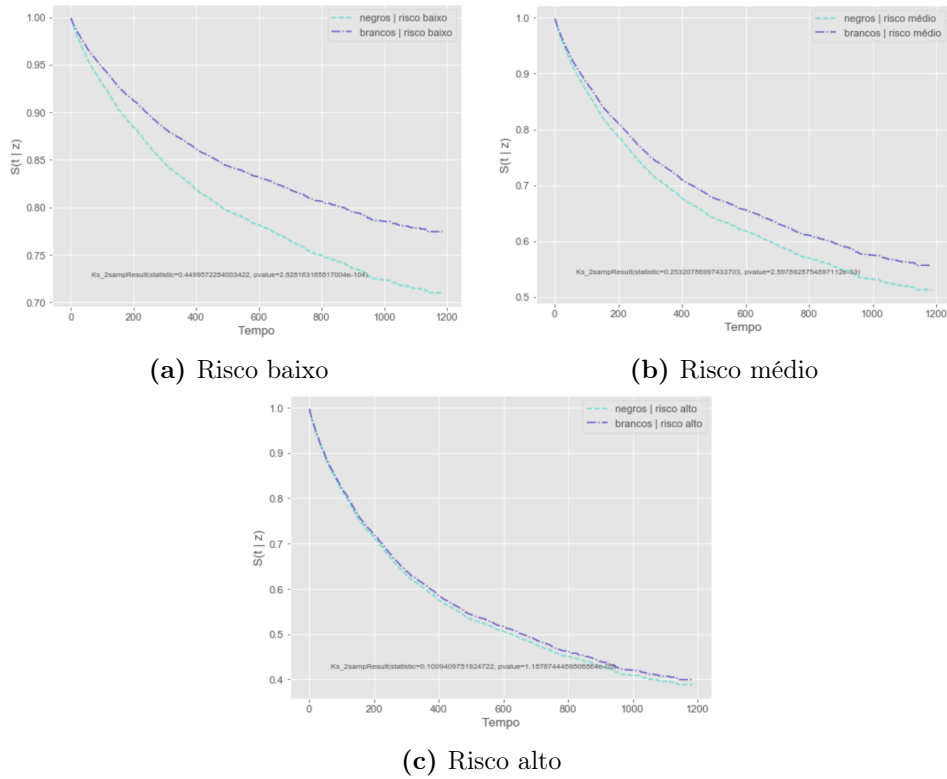


Figura 6.15. Curvas previstas para a proposta 2 na base COMPAS

Tabela 6.17. Resultados do teste KS para proposta 2 na base COMPAS

Grupos	KS	Valor-p
Negros risco baixo / Brancos risco baixo	0.45	$2.52e^{-104}$
Negros risco médio / Brancos risco médio	0.25	$2.60e^{-33}$
Negros risco alto / Brancos risco alto	0.10	$1.15e^{-05}$

estão de acordo. Logo, esses casos satisfazem a proposta de divergência em paridade demográfica condicionada. Os resultados resumidos encontram-se na Tabela 6.18.

Tabela 6.18. Resultados para a proposta de divergência em paridade demográfica condicionada na base COMPAS

Grupos	Resultado
Negros risco baixo / Brancos risco baixo	Justo
Negros risco médio / Brancos risco médio	Justo
Negros risco alto / Brancos risco alto	Justo

6.3 Proposta 3: Discriminação causal

A terceira proposta, vista em 4.1.3, é a utilização do erro dado pelo c-index para comparação dos resultados entre indivíduos similares que diferem apenas na variável sensível z , tal que $erro(z_1) = erro(z_2)$, onde z_1, z_2 são grupos com variáveis sensíveis diferentes entre si.

6.3.1 MIMIC-III

Foram usados oito grupos (mulheres, homens, negros, brancos, mulheres negras, mulheres brancas, homens negros, homens brancos). Dado a limitação de pessoas com características muito similares a ponto de só diferenciar a variável sensível, a solução proposta foi gerar indivíduos idênticos com variáveis sensíveis opostas e calcular o erro para cada grupo.

A Tabela 6.19 exibe os resultados para cada um dos grupos testados. Todos os casos foram considerados justos e portanto satisfazem a definição de discriminação causal.

O experimento foi repetido 1000x usando uma técnica chamada Bootstrap (Efron, 1979) com repetição e um intervalo de confiança (IC) de 95%. O erro foi calculado para o grupo original e depois comparado com o do grupo criado artificialmente. Dado que a única diferença entre os grupos é em relação a variável sensível, foi considerada injusta uma diferença maior que 0.05 em relação ao resultado de cada grupo. Os demais casos foram considerados justos. Esse limiar de 0.05 foi definido levando em consideração que mudanças dessa escala em relação ao c-index são significativas.

Tabela 6.19. Resultados para a proposta de discriminação causal

Erro grupo original - IC	Erro grupo artificial - IC	Resultado
mulheres - (0.284, 0.287)	homens - (0.284, 0.287)	Justo
homens - (0.291, 0.293)	mulheres - (0.291, 0.293)	Justo
negros - (0.327, 0.333)	brancos - (0.327, 0.333)	Justo
brancos - (0.275, 0.276)	negros - (0.275, 0.276)	Justo
mulheres negras - (0.371, 0.379)	mulheres brancas - (0.371, 0.379)	Justo
mulheres brancas - (0.273, 0.275)	mulheres negras - (0.273, 0.275)	Justo
homens negros - (0.257, 0.263)	homens brancos - (0.257, 0.263)	Justo
homens brancos - (0.278, 0.281)	homens negros - (0.278, 0.281)	Justo

6.3.2 Rossi

Para a terceira proposta, foram usados seis grupos levando em consideração raça, estado civil e financiamento. Novamente, dado a limitação de pessoas com características muito similares a ponto de só diferenciar a variável sensível, a solução foi gerar indivíduos idênticos com variáveis sensíveis opostas e calcular o erro para cada grupo.

A Tabela 6.20 exibe os resultados para cada um dos grupos testados. O experimento foi repetido 1000x usando bootstrap com repetição e o intervalo de confiança (IC) do resultado é de 95%. O erro foi calculado para o grupo original e depois comparado com o do grupo criado artificialmente. Dado que a única diferença entre os grupos é em relação a variável sensível, foi considerada injusta uma diferença maior que 0.05 em relação ao resultado de cada grupo. Os demais casos foram considerados justos.

Tabela 6.20. Resultados para a proposta de discriminação causal para a base Rossi

Erro grupo original - IC	Erro grupo artificial - IC	Resultado
negros - (0.369, 0.372)	outros - (0.371, 0.374)	Justo
outros - (0.310, 0.320)	negros - (0.309, 0.319)	Justo
casados - (0.312, 0.327)	não casados - (0.316, 0.332)	Justo
não casados - (0.386, 0.389)	casados - (0.386, 0.390)	Justo
financiados - (0.401, 0.406)	não financiados - (0.358, 0.362)	Justo
não financiados - (0.359, 0.363)	financiados - (0.402, 0.407)	Justo

Todos os casos presentes na Tabela 6.20 foram considerados justos e portanto satisfazem a definição de discriminação causal.

6.3.3 COMPAS

Para a terceira proposta, foram usados oito grupos levando em consideração raça e fator de risco. Novamente, dado a limitação de pessoas com características muito similares a ponto de só diferenciar a variável sensível, a solução foi gerar indivíduos idênticos com variáveis sensíveis opostas e calcular o erro para cada grupo.

A Tabela 6.21 exibe os resultados para cada um dos grupos testados. Como todos os casos foram considerados justos, todos satisfazem a definição de discriminação causal. O experimento foi repetido 100x usando bootstrap com repetição e o intervalo de confiança (IC) do resultado é de 95%. A quantidade de repetições foi diferente das anteriores devido a uma limitação de recursos computacionais e tempo para execução.

O erro foi calculado para o grupo original e depois comparado com o do grupo criado artificialmente. Dado que a única diferença entre os grupos é em relação a variável sensível, foi considerada injusta uma diferença maior que 0.05 em relação ao resultado de cada grupo. Os demais casos foram considerados justos.

Tabela 6.21. Resultados para a proposta de discriminação causal para a base COMPAS

Erro grupo original - IC	Erro grupo artificial - IC	Resultado
negros - (0.382, 0.384)	brancos - (0.378, 0.382)	Justo
brancos - (0.377, 0.381)	negros - (0.382, 0.385)	Justo
negros risco baixo - (0.499, 0.499)	brancos risco baixo - (0.499, 0.499)	Justo
brancos risco baixo - (0.499, 0.499)	negros risco baixo - (0.499, 0.499)	Justo
negros risco médio - (0.498, 0.499)	brancos risco médio - (0.498, 0.498)	Justo
brancos risco médio - (0.498, 0.498)	negros risco médio - (0.498, 0.499)	Justo
negros risco alto - (0.498, 0.498)	brancos risco alto - (0.496, 0.496)	Justo
brancos risco alto - (0.496, 0.496)	negros risco alto - (0.498, 0.498)	Justo

6.4 Proposta 4: Justiça de filas

Para a última proposta, vista em 4.1.4, foi implementada uma nova função dentro da biblioteca Pycox para o cálculo do c-index. Usamos como base a função *concordance_td* e fizemos as alterações necessárias para permitir a comparação entre grupos com variáveis sensíveis diferentes e não mais pares aleatórios de indivíduos. Essa versão modificada está disponível em <https://github.com/bseewald/pycox>.

6.4.1 MIMIC-III

Para esta base o erro representado por $erro(z_1, z_2)$ favorece o grupo com variável z_2 e o $erro(z_2, z_1)$ favorece o grupo com variável z_1 . No primeiro erro a predição disse que a pessoa com variável z_1 iria morrer depois que a pessoa com variável z_2 porém ele morreu antes, ou seja favorecendo z_2 e prejudicando z_1 . O segundo erro é complementar a esse primeiro. Em ambos os casos, isso significa que foram dadas prioridades de atendimento de forma incorreta.

A Tabela 6.22 apresenta os resultados de 1000 experimentos usando bootstrap com repetição, com intervalo de confiança (IC) de 95%. Foram considerados casos justos aqueles que ambos os erros tiveram diferença igual ou menor que 0.05 e casos injustos os erros que tiveram diferença maior que 0.05.

Tabela 6.22. Resultados para a quarta proposta

Grupos: $z_1 - z_2$	$Erro(z_1, z_2)$ IC	- $Erro(z_2, z_1)$ IC	- Resultado
Mulheres - homens	(0.524, 0.526)	(0.622, 0.624)	Injusto para ho- mens
Negros - brancos	(0.599, 0.604)	(0.590, 0.593)	Justo
Mulheres negras - mu- lheres brancas	(0.626, 0.633)	(0.607, 0.611)	Justo
Homens negros - ho- mens brancos	(0.561, 0.566)	(0.555, 0.561)	Justo

Dos quatro casos, somente um foi considerado injusto: mulheres / homens, onde a diferença do erro foi de 0.1 ($|0.52 - 0.62|$). Isto significa que o erro beneficiou mais o grupo z_1 , mulheres, uma vez que a prioridade de atendimento de forma incorreta aconteceu em mais ocasiões do que para o outro grupo.

6.4.2 Rossi

Foram usados os mesmos grupos da proposta de discriminação causal e a função criada para a biblioteca Pycox foi utilizada aqui para cálculo da métrica. Para esta base, a proposta do erro representado por $erro(z_1, z_2)$ desfavorece o grupo com variável z_2 e o $erro(z_2, z_1)$ desfavorece o grupo com variável z_1 . Aqui, o $erro(z_1, z_2)$ diz que a predição de reincidência do z_1 iria ocorrer depois que a pessoa com variável z_2 porém ele reincidiu antes, ou seja, isso favoreceu z_1 e prejudicou z_2 . O $erro(z_2, z_1)$ possui explicação complementar a anterior. Em ambos os casos, isso significa que houve uma previsão de reincidência de forma incorreta.

Importante notar que esse proposta depende da base que está sendo utilizada, pois dependendo da situação o grupo favorecido pelo erro irá mudar. No caso da base MIMIC-III, o $erro(z_1, z_2)$ favorece o grupo com variável z_2 e aqui favorece o grupo com variável z_1 . Caso esse detalhe não seja percebido, pode originar um resultado errado.

A Tabela 6.23 apresenta os resultados de 1000 experimentos usando bootstrap com repetição e intervalo de confiança (IC) de 95%. Foram considerados casos justos aqueles que ambos os erros tiveram diferença igual ou abaixo de 0.05 e casos injustos os erros que tiveram diferença maior que 0.05.

Tabela 6.23. Resultados para a quarta proposta junto a base Rossi

Grupos: $z_1 - z_2$	$Erro(z_1, z_2)$ - IC	$Erro(z_2, z_1)$ - IC	Resultado
negros - outros	(0.715, 0.719)	(0.727, 0.736)	Justo
casados - não casados	(0.828, 0.838)	(0.506, 0.512)	Injusto para não casados
financiados - não fi- nanciados	(0.862, 0.866)	(0.632, 0.637)	Injusto para não financi- ados

Dos três casos, dois foram considerados injustos: casados / não casados e financiados / não financiados. No primeiro caso a diferença do erro foi de 0.32 ($|0.83 - 0.51|$) e no segundo a diferença do erro foi de 0.23 ($|0.86 - 0.63|$). Em ambos os casos isto significa que o erro prejudicou mais o grupo z_2 (não casados / não financiados) uma vez que a reincidência prevista pelo modelo irá acontecer antes para esses grupos.

6.4.3 COMPAS

Foram usados os mesmos grupos da proposta de discriminação causal. Para esta base, a proposta do erro representado por $erro(z_1, z_2)$ desfavorece o grupo com variável z_2 e o $erro(z_2, z_1)$ desfavorece o grupo com variável z_1 . Aqui, o $erro(z_1, z_2)$ diz que a predição de reincidência do z_1 iria ocorrer depois que a pessoa com variável z_2 porém ele reincidiu antes, ou seja, isso favoreceu z_1 e prejudicou z_2 . O $erro(z_2, z_1)$ possui explicação complementar a anterior. Em ambos os casos, isso significa que foi prevista uma reincidência de forma incorreta.

A Tabela 6.24 apresenta os resultados de 100 experimentos usando bootstrap com repetição e o intervalo de confiança é de 95%. Foram considerados casos justos aqueles que ambos os erros tiveram diferença igual ou menor que 0.05 e casos injustos os erros que tiveram diferença maior que 0.05.

Tabela 6.24. Resultados para a quarta proposta junto a base COMPAS

Grupos: $z_1 - z_2$	$Erro(z_1, z_2)$ - IC	$Erro(z_2, z_1)$ - IC	Resultado
negros - brancos	(0.665, 0.669)	(0.911, 0.913)	Injusto para ne- gros
negros risco baixo - brancos risco baixo	(0.807, 0.812)	(0.998, 0.998)	Injusto para ne- gros
negros risco médio - brancos risco médio	(0.929, 0.931)	(0.998, 0.998)	Injusto para ne- gros
negros risco alto - brancos risco alto	(0.983, 0.983)	(0.997, 0.997)	Justo

Dos quatro casos, três foram considerados injustos. No primeiro caso a diferença do erro foi de 0.25 ($|0.91 - 0.66|$), no segundo a diferença do erro foi de 0.189 ($|0.99 - 0.81|$) e no terceiro caso a diferença do erro foi de 0.06 ($|0.99 - 0.93|$). Em todos os casos isto significa que o erro prejudicou mais o grupo z_1 , pessoas negras, uma vez que a reincidência prevista pelo modelo irá acontecer antes para esses grupos.

Neste capítulo foram apresentados e discutidos os resultados das quatro propostas sugeridas nesta dissertação. Para a base MIMIC-III foram encontrados vieses na proposta de divergência em paridade demográfica, divergência em paridade demográfica condicionada além de um caso de injustiça na avaliação da métrica de justiça de

filas. Para a base Rossi houveram dois casos de injusta na métrica de justiça de filas enquanto na base COMPAS apareceram 3 situações de injusta para a métrica de justiça de filas.

Essas ocorrências de injustiça em bases como MIMIC-III, Rossi e COMPAS envolvendo análise de sobrevivência demonstram as possibilidades na área, desde o estudo de novas bases até propostas de mitigação do problema. No Capítulo 7 iremos aprofundar essa discussão e apontar caminhos para trabalhos futuros.

Capítulo 7

Conclusão

Essa dissertação apresentou 3 possíveis abordagens de aplicabilidade de conceitos de justiça em modelos de análise sobrevivência. A primeira focou na disparidade das curvas de sobrevivência observadas nos dados quando comparadas com previsões, denominada de divergência em paridade demográfica. Para isso, utilizamos o método de Kaplan-Meier para as curvas empíricas e o modelo de Cox para as curvas preditas. A segunda abordagem, denominada discriminação causal, consistiu na realização de um cálculo do c-index, no qual alteramos nos dados o grupo de interesse. Por fim, propusemos uma métrica nova chamada de justiça de filas, na qual comparamos cenários hipotéticos de duas pessoas sendo julgadas por um modelo de aprendizado de máquina ao mesmo tempo.

Essas abordagens foram testadas em 3 bases de dados com contextos diferentes, a saber MIMIC-III, Rossi e COMPAS, a primeira sendo uma base médica e as demais criminais, com o intuito de analisar justiça sob diferentes aspectos e que foram discutidos no Capítulo 3. Além disso, foram testados sete algoritmos de aprendizado de máquina (Cox, CoxNet, RSF, Cox-MLP, Cox-PH, Cox-Time e DeepHit), tendo sido o algoritmo Cox-Time o escolhido para fazer as previsões e os cálculos das métricas envolvendo o c-index.

Para a base MIMIC-III, foram encontrados vieses nas propostas de divergência em paridade demográfica, divergência em paridade demográfica condicionada e justiça de filas. No entanto, no caso da primeira proposta, não fica nítido se o viés beneficia pessoas negras ou brancas. No segundo caso, o viés aparece em ambas as raças e no terceiro caso, o viés aparece beneficiando mulheres. Esses resultados, no entanto, precisam ser avaliados com cuidado, pois se trata de uma base extensa, que apresenta complexidades inerentes da área da saúde e possui diversas outras variáveis que poderiam ser trabalhadas mas que não foram alvo dessa dissertação.

Para a base Rossi não foram encontrados vieses nas propostas de divergência em paridade demográfica, divergência em paridade demográfica condicionada e discriminação causal. No entanto, a métrica de justiça de filas identificou dois casos de injustiça, no qual o erro prejudicou os grupos de pessoas não casadas e o grupo de pessoas não financiadas, dada que a reincidência prevista pelo modelo irá acontecer antes para esses grupos. Importante salientar que apesar de não ter sido encontrado viés quando considerada a variável raça, a análise exploratória mostrou que ela influencia o risco de ocorrência do evento, neste caso, reincidência. Portanto, é importante o estudo de outras definições de justiça aplicadas a análise de sobrevivência que possam captar esse impacto.

Por fim, na base COMPAS, também não foram encontrados vieses nas propostas de divergência em paridade demográfica, divergência em paridade demográfica condicionada e discriminação causal. Para a métrica de justiça de filas, foram identificados casos de injustiça em 3 das 4 situações propostas. Nesses casos, o erro prejudicou mais o grupo de pessoas negras, uma vez que a reincidência prevista pelo modelo irá acontecer antes para esses grupos. Esse resultado está em concordância com os achados da análise original, que também identificou vieses nessa base.

Em geral, os resultados para divergência em paridade demográfica condicionada na base MIMIC-III e os resultados da métrica de justiça de filas, em todas as bases, se mostraram satisfatórios. Em todos os casos, apareceram situações com vieses, o que mostra que as abordagens propostas são um caminho para a aplicabilidade de conceitos de justiça em modelos de análise sobrevivência. Em particular, a métrica de justiça de filas mostrou-se bem promissora, conseguindo identificar casos de injustiça em todas as bases selecionadas. Além disso, no que tange ao uso de bases da área médica, encontramos diferenças que precisam ser melhor exploradas, dada que ainda não é totalmente compreensível como pesquisadores conseguem quantificar justiça em serviços de saúde. Também é essencial garantir que essas bases possuam dados mais diversos e representativos, para que as comparações sejam mais robustas e confiáveis. Este tipo de avaliação é crucial para evitar que as ferramentas que utilizam modelos de aprendizado de máquina perpetuem injustiças sociais e históricas.

Como sugestão de trabalhos futuros, pode-se testar bases que tenham outros contextos, por exemplo bases de empréstimo bancário e contratação de pessoas, assim ampliando o conjunto de resultados. Pode-se também utilizar um tempo fixo nos dados e verificar se existe impacto sobre os resultados atuais. Há espaço para melhorias na própria métrica de justiça de filas, focando em facilitar o entendimento da mesma. De forma mais avançada, é possível aplicar técnicas para mitigar os vieses encontrados nessas bases, podendo até virar uma ferramenta nos mesmo moldes dos projetos

apresentados na Tabela 3.3.

Em particular, para a base MIMIC-III seria interessante avaliar o uso de outras variáveis, principalmente as clínicas, para entender como elas se relacionam com justiça. Há a parte de prontuários médicos em que poderiam ser aplicadas técnicas de processamento de linguagem natural para extrair novas variáveis e obter mais dados sobre cada paciente.

Em suma, esse trabalho mostrou que ainda há muitas oportunidades para pesquisa e possíveis inovações para a área.

Referências Bibliográficas

- Abebe, R.; Barocas, S.; Kleinberg, J.; Levy, K.; Raghavan, M. & Robinson, D. G. (2020). Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 252--260.
- Antolini, L.; Boracchi, P. & Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927--3944.
- Barocas, S. & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104:671.
- Barry-Jester, A. M.; Casselman, B. & Goldstein, D. (2015). The new science of sentencing. *The Marshall Project*, 4:2015.
- Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A. et al. (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.
- Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M. & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, p. 0049124118782533.
- Beutel, A.; Chen, J.; Doshi, T.; Qian, H.; Woodruff, A.; Luu, C.; Kreitmann, P.; Bischof, J. & Chi, E. H. (2019). Putting fairness principles into practice: Challenges, metrics, and improvements. *arXiv preprint arXiv:1901.04562*.
- Caton, S. & Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.
- Char, D. S.; Shah, N. H. & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine*, 378(11):981.

- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153--163.
- Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S. & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797--806.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187--202.
- Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269--276.
- Cutler, S. J. & Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival. *Journal of chronic diseases*, 8(6):699--712.
- Diakopoulos, N.; Friedler, S.; Arenas, M.; Barocas, S.; Hay, M.; Howe, B.; Jagadish, H.; Unsworth, K.; Sahuguet, A.; Venkatasubramanian, S. et al. (2017). Principles for accountable algorithms and a social impact statement for algorithms. *FAT/ML*.
- Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O. & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214--226.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1--26.
- Friedman, J.; Hastie, T.; Tibshirani, R. et al. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.
- Gajane, P. & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*.
- Galhotra, S.; Brun, Y. & Meliou, A. (2017). Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pp. 498--510.
- Glazier, S. W. (2019). Sequential survival analysis with deep learning.
- Goel, N.; Yaghini, M. & Faltings, B. (2018). Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

- Goodman, S. N.; Goel, S. & Cullen, M. R. (2018). Machine learning, health disparities, and causal reasoning. *Annals of internal medicine*.
- Graf, E.; Schmoor, C.; Sauerbrei, W. & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529--2545.
- Grgic-Hlaca, N.; Redmiles, E. M.; Gummadi, K. P. & Weller, A. (2018). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*, pp. 903--912.
- Grgic-Hlaca, N.; Zafar, M. B.; Gummadi, K. P. & Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, p. 2.
- Gummadi, K. P. (2019). Foundations for fair algorithmic decision making. In *EGC*, pp. 3--4.
- Guo, C. & Berkhahn, F. (2016). Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*.
- Hardt, M.; Price, E. & Srebro, N. (2016). Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*.
- Harrell, F. E.; Califf, R. M.; Pryor, D. B.; Lee, K. L. & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, 247(18):2543--2546.
- Heagerty, P. J. & Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92--105.
- Hodges, J. L. (1958). The significance probability of the smirnov two-sample test. *Arkiv för Matematik*, 3(5):469--486.
- Ishwaran, H.; Kogalur, U. B.; Blackstone, E. H.; Lauer, M. S. et al. (2008). Random survival forests. *Annals of Applied Statistics*, 2(3):841--860.
- Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A. & Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.
- Johnson, A. E. W.; Kramer, A. A. & Clifford, G. D. (2013). A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. *Critical care medicine*, 41(7):1711--1718.

- Kamishima, T.; Akaho, S.; Asoh, H. & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35--50. Springer.
- Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457--481.
- Katzman, J. L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T. & Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1--12.
- Kirkpatrick, K. (2017). It's not the algorithm, it's the data. *Communications of the ACM*, 60(2):21--23.
- Klein, J. P. & Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.
- Kleinberg, J.; Mullainathan, S. & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
- Krasanakis, E.; Spyromitros-Xioufis, E.; Papadopoulos, S. & Kompatsiaris, Y. (2018). Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference*, pp. 853--862.
- Kusner, M. J.; Loftus, J. R.; Russell, C. & Silva, R. (2017). Counterfactual fairness. *arXiv preprint arXiv:1703.06856*.
- Kvamme, H.; Borgan, Ø. & Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of machine learning research*, 20(129):1--30.
- Lee, C.; Zame, W.; Yoon, J. & van der Schaar, M. (2018). Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Lee, E. T. & Wang, J. (2003). *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons.
- Loshchilov, I. & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K. & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

- Menon, A. K. & Williamson, R. C. (2018). The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pp. 107--118. PMLR.
- Miller, C. C. (2015). Can an algorithm hire better than a human. *The New York Times*, 25.
- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
- Narayanan, A. (2018). Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*.
- Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825--2830.
- Petrasic, K.; Saul, B.; Greig, J.; Bornfreund, M. & Lamberth, K. (2017). Algorithms and bias: What lenders need to know. *White & Case*.
- Pirracchio, R.; Petersen, M. L.; Carone, M.; Rigon, M. R.; Chevret, S. & van der Laan, M. J. (2015). Mortality prediction in intensive care units with the super icu learner algorithm (sicula): a population-based study. *The Lancet Respiratory Medicine*, 3(1):42--52.
- Powell, J. L. (1994). Estimation of semiparametric models. *Handbook of econometrics*, 4:2443--2521.
- Rajkomar, A.; Hardt, M.; Howell, M. D.; Corrado, G. & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866--872.
- Reddy, C. K.; Li, Y. & Aggarwal, C. (2015). A review of clinical prediction models. *Healthcare data analytics*, 36:343--378.
- Rossi, P. H.; Berk, R. A. & Lenihan, K. J. (1980). Money, work and crime: some experimental results.
- Russell, S. J. & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.
- Saleiro, P.; Kuester, B.; Hinkson, L.; London, J.; Stevens, A.; Anisfeld, A.; Rodolfa, K. T. & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.

- Saxena, N.; Huang, K.; DeFilippis, E.; Radanovic, G.; Parkes, D. & Liu, Y. (2018). How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. *arXiv preprint arXiv:1811.03654*.
- Verma, S. & Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1--7. IEEE.
- Wang, P.; Li, Y. & Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1--36.
- Wu, Y.; Zhang, L. & Wu, X. (2018). Fairness-aware classification: Criterion, convexity, and bounds. *arXiv preprint arXiv:1809.04737*.
- Xie, F.; Chakraborty, B.; Ong, M. E. H.; Goldstein, B. A. & Liu, N. (2020). Autoscore: A machine learning-based automatic clinical score generator and its application to mortality prediction using electronic health records. *JMIR medical informatics*, 8(10):e21798.
- Zafar, M. B.; Valera, I.; Gomez Rodriguez, M. & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171--1180. International World Wide Web Conferences Steering Committee.

Apêndice A

Resultados complementares

Neste capítulo serão disponibilizados gráficos e tabelas complementares as discussões sobre divergência em paridade demográfica condicionada para a base MIMIC-III, que estão presentes no Capítulo 6.

O método de Kaplan-Meier (KM) é uma observação da função de sobrevivência de cada grupo a partir dos dados coletados enquanto o teste estatístico LogRank compara as probabilidades de sobrevivência de cada curva entre os diferentes grupos. Relembrando quais testes de hipótese foram feitos, temos que a hipótese nula é que não há diferença entre os grupos e a hipótese alternativa é que há diferença entre os grupos. Em seguida, com o uso do modelo de Cox treinado, fizemos a predição da curva de sobrevivência para cada um desses mesmos grupos e também aplicamos o teste estatístico de Kolmogorov-Smirnov (KS) (Hodges, 1958). Para esse teste a hipótese nula é que as distribuições são idênticas e a hipótese alternativa é que elas não são idênticas.

Para essa proposta criamos quarenta grupos de teste dividindo entre gênero, raça e gênero-raça. Os fatores condicionais escolhidos foram a pontuação Oasis e quatro tipos de diagnóstico dados pelo hospital: câncer, diabetes, coração e transplante.

- **Kaplan-Meier**

A curva KM para cada um desses grupos é mostrada nas Figuras A.1, A.2, A.3 e A.4. O tamanho da amostra para cada um desses grupos é diferente, por isso há curvas com tempo de sobrevivência menores que outras, dado que essas observações são limitadas. Nas curvas da Figura A.1, que diferem pelo gênero e fatores condicionais, a curva de transplante tem uma diferença visível a partir de 75 dias de internação, com mulheres tendo melhores probabilidades de sobrevivência do que homens. Na Figura A.2 é possível ver a limitação nos pontos observados, no entanto o gráfico (b) mostra que a curva de sobrevivência

de pessoas negras com o fator condicional câncer é pior do que para pessoas brancas. Já na Figura A.3, o gráfico (e) mostra curvas bem distintas, no entanto o tamanho da amostra deve ser considerada nesse caso. Por fim, o gráfico (b) da Figura A.4 também mostra curvas que após um determinado tempo começam a ter comportamentos diferentes, sendo a curva de sobrevivência pior para homens negros com câncer do que para homens brancos com o mesmo fator condicional. Os resultados do teste LogRank estão presentes na Tabela A.1. Para todos os casos citados o valor-p é menor que 5%, ou seja, rejeitando a hipótese nula. Desta forma, pode-se concluir que há diferença significativa entre as curvas.

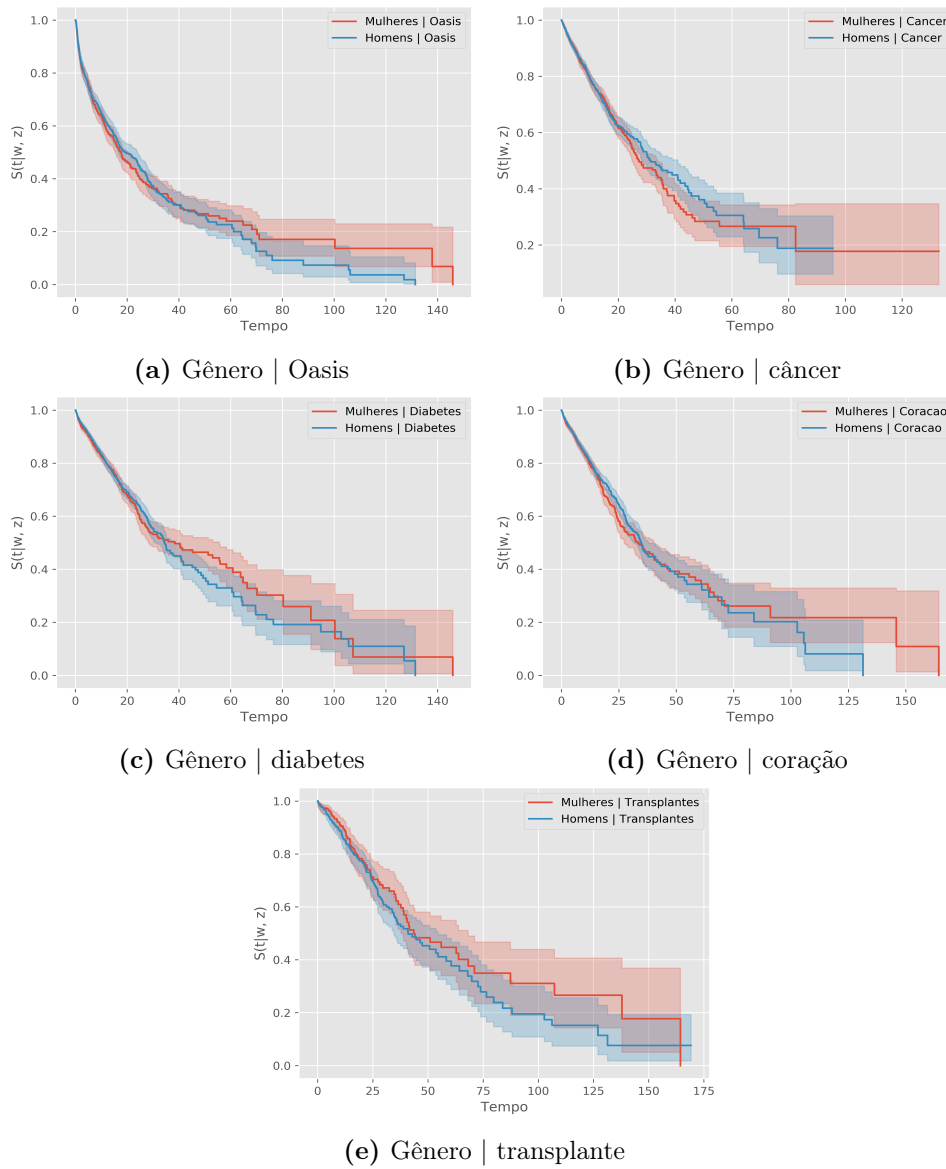


Figura A.1. Curvas KM com base no gênero para a proposta 2 na base MIMIC-III

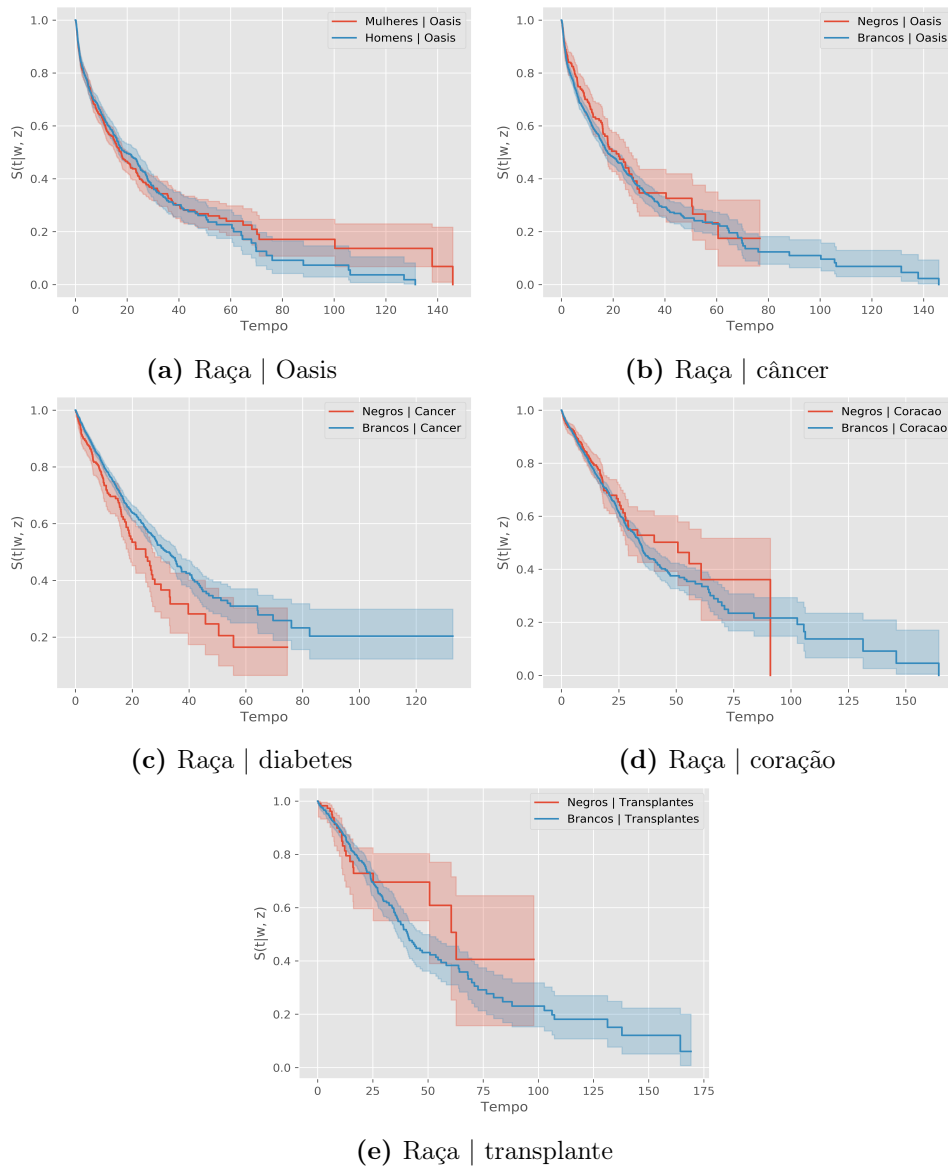


Figura A.2. Curvas KM com base na raça para a proposta 2 na base MIMIC-III

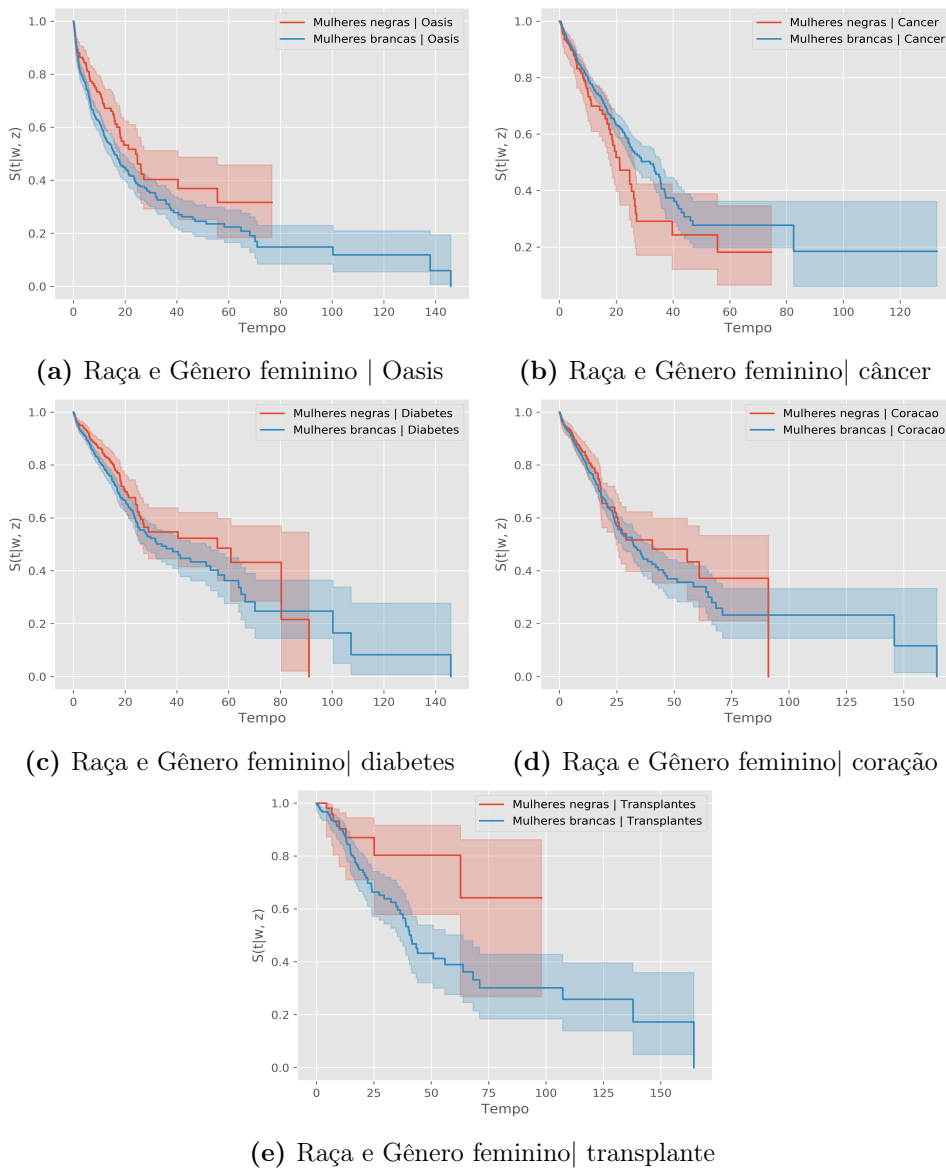


Figura A.3. Curvas KM com base na raça e gênero feminino para a proposta 2 na base MIMIC-III

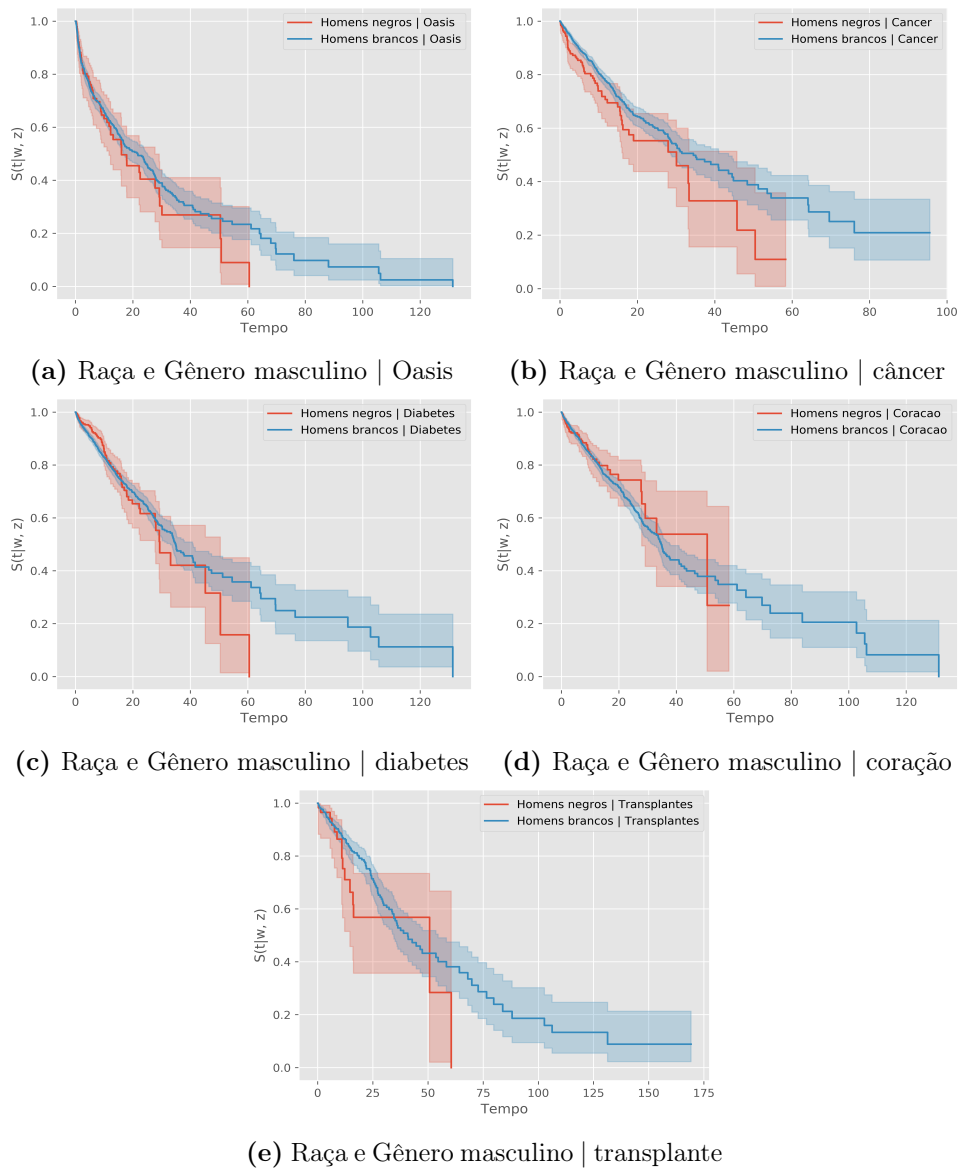


Figura A.4. Curvas KM com base na raça e gênero masculino para a proposta 2 na base MIMIC-III

Tabela A.1. Tabela com resultados do teste LogRank para a proposta 2 da base MIMIC-III

Grupos	LR	Valor-p
Mulheres / Homens Oasis	0.078	0.778
Mulheres / Homens câncer	0.38	0.53
Mulheres / Homens diabetes	0.07	0.78
Mulheres / Homens coração	0.67	0.41
Mulheres / Homens transplante	1.94	0.16
Negros / Brancos Oasis	1.83	0.17
Negros / Brancos câncer	9.8	0.00017
Negros / Brancos diabetes	2.97	0.08
Negros / Brancos coração	1.07	0.3
Negros / Brancos transplante	0.62	0.43
Mulheres negras / mulheres brancas Oasis	6.24	0.012
Mulheres negras / mulheres brancas câncer	3.81	0.05
Mulheres negras / mulheres brancas diabetes	4.5	0.03
Mulheres negras / mulheres brancas coração	0.91	0.33
Mulheres negras / mulheres brancas transplante	3.8	0.051
Homens negros / homens brancos Oasis	0.75	0.38
Homens negros / homens brancos câncer	5.83	0.015
Homens negros / homens brancos diabetes	0.05	0.82
Homens negros / homens brancos coração	0.54	0.45
Homens negros / homens brancos transplante	1.19	0.27

- **Curvas preditas com o modelo de Cox**

Assim como nas curvas KM, avaliamos os resultados das curvas preditas. Além dos casos apresentados no Capítulo 6 que apresentaram disparidade entre as curvas KM e preditas, os demais casos não tiveram disparidade. Todos os casos são apresentados nas Figuras A.5, A.6, A.7 e A.8. Os resultados do teste estatístico Kolmogorov-Smirnov (KS) encontram-se na Tabela A.2. E o resumo dos resultados da comparação entre as curvas encontra-se no Capítulo 6 na Tabela 6.12.

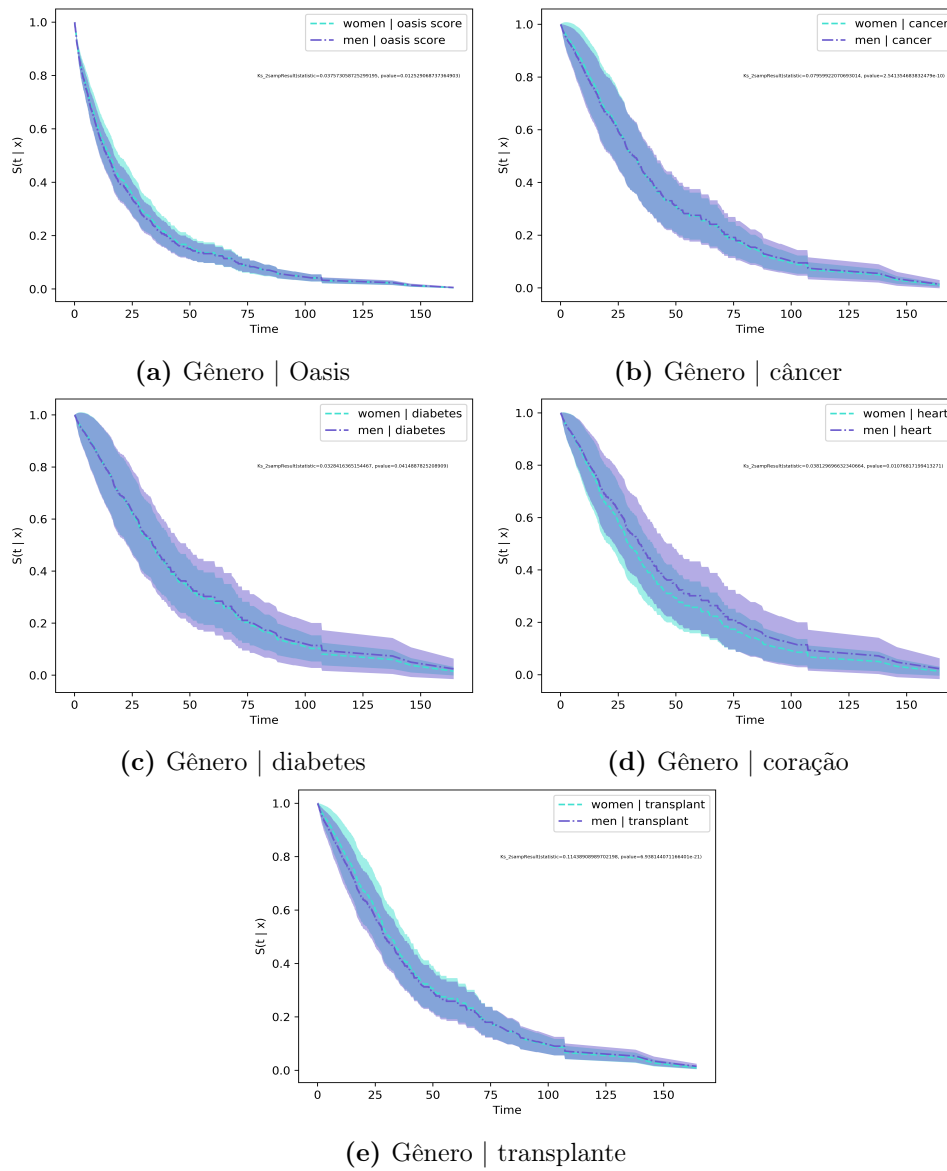


Figura A.5. Curvas preditas com base no gênero para a proposta 2 na base MIMIC-III

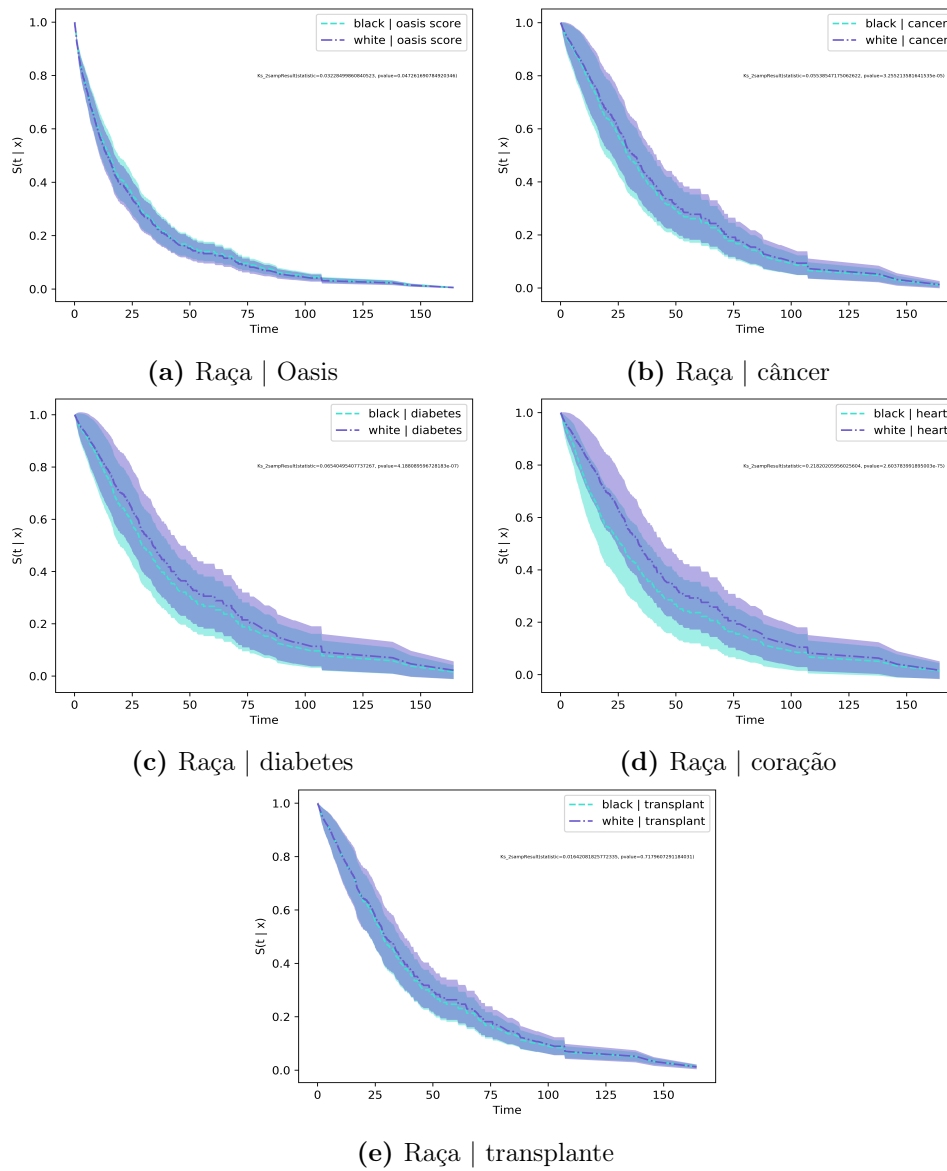


Figura A.6. Curvas preditas com base na raça para a proposta 2 na base MIMIC-III

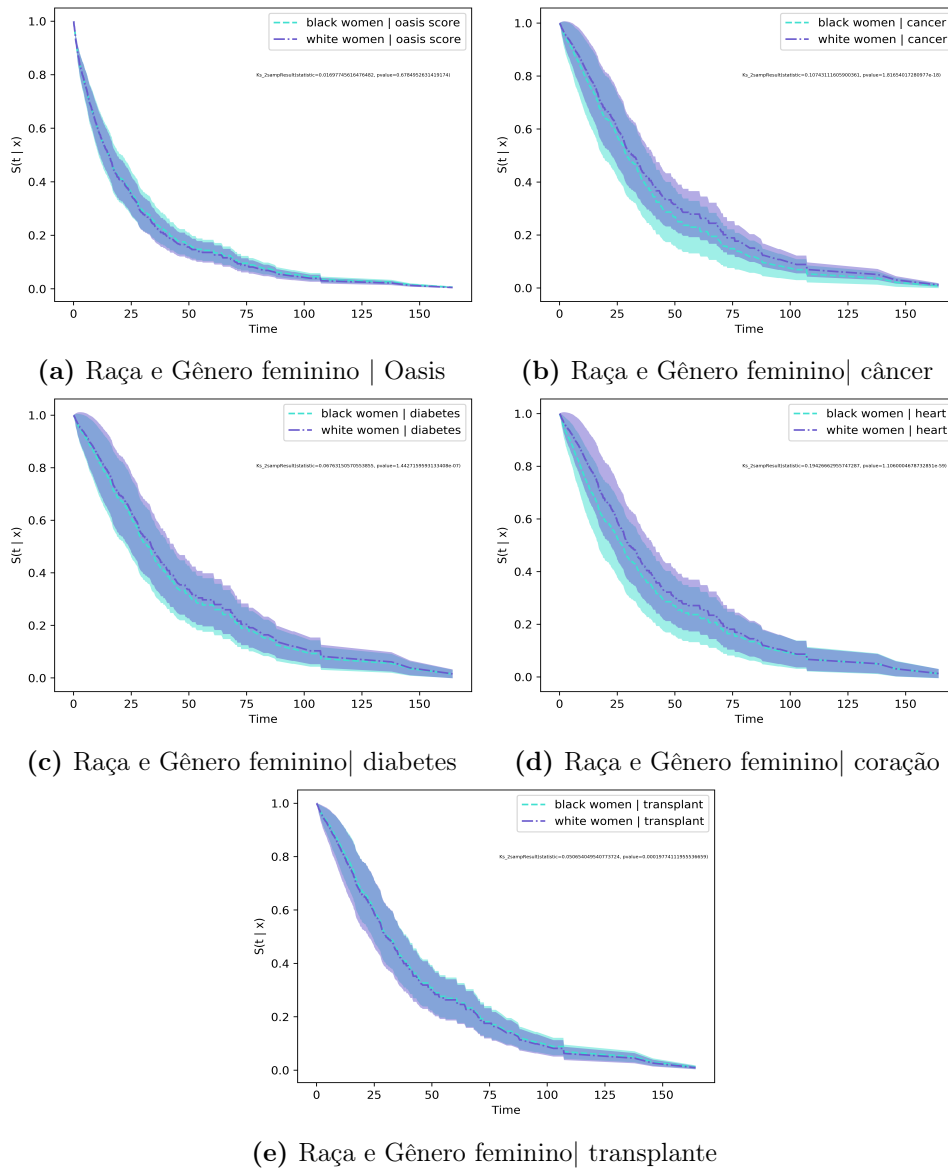


Figura A.7. Curvas previstas com base na raça e gênero feminino para a proposta 2 na base MIMIC-III

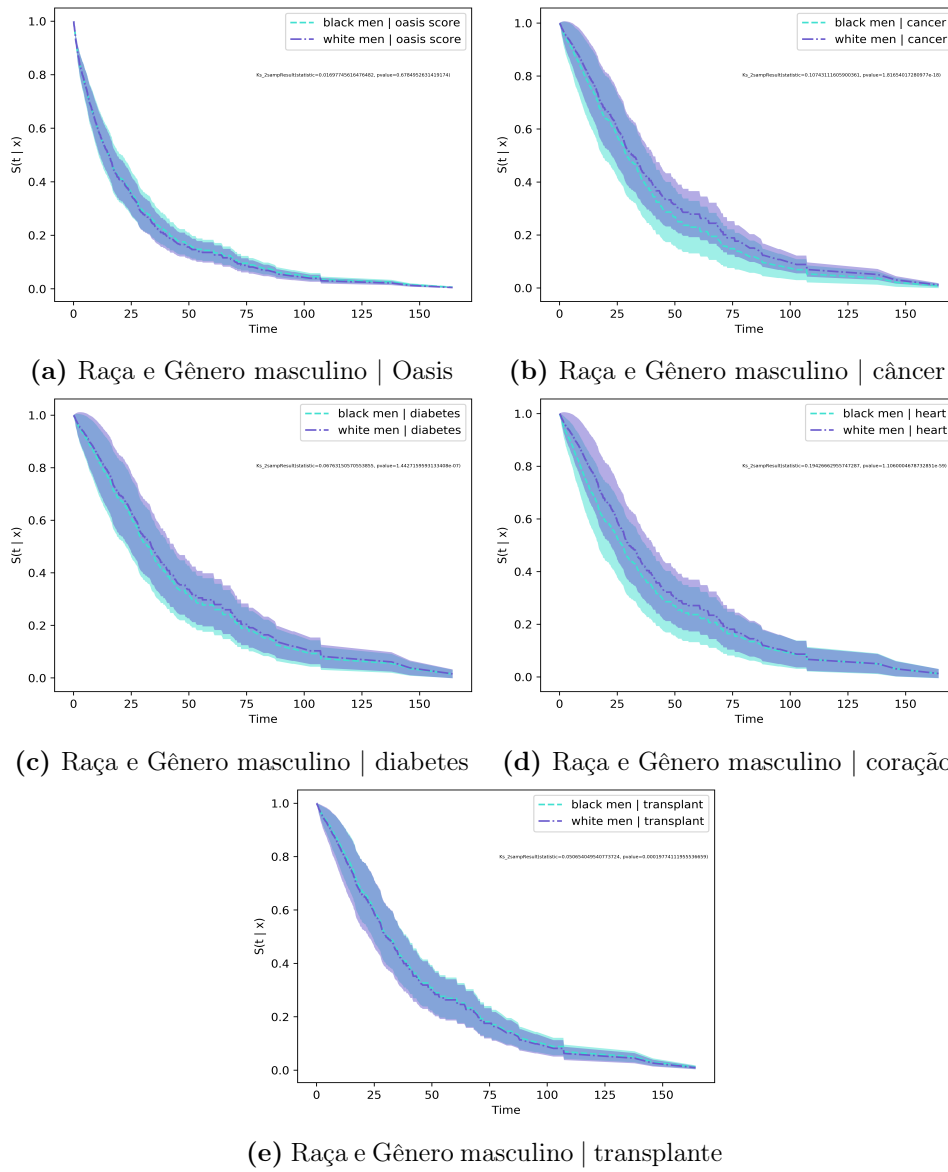


Figura A.8. Curvas previstas com base na raça e gênero masculino para a proposta 2 na base MIMIC-III

Tabela A.2. Resultados do teste KS para a proposta 2 da base MIMIC-III

Grupos	KS	Valor-p
Mulheres Oasis / Homens Oasis	0.02	0.365
Mulheres câncer / Homens câncer	0.08	$2.54e^{-10}$
Mulheres coração / Homens coração	0.10	$2.32e^{-17}$
Mulheres diabetes / Homens diabetes	0.04	0.013
Mulheres transplante / Homens transplante	0.01	0.79
Negros Oasis / Brancos Oasis	0.01	0.98
Negros câncer / Brancos câncer	0.12	$2.37e^{-23}$
Negros coração / Brancos coração	0.23	$8.19e^{-83}$
Negros diabetes / Brancos diabetes	0.12	$2.63e^{-22}$
Negros transplante / Brancos transplante	0.04	0.004
Mulheres negras Oasis / mulheres brancas Oasis	0.04	0.003
Mulheres negras câncer / mulheres brancas câncer	0.15	$1.69e^{-34}$
Mulheres negras coração / mulheres brancas coração	0.21	$3.90e^{-71}$
Mulheres negras diabetes / mulheres brancas diabetes	0.11	$2.25e^{-18}$
Mulheres negras transplante / mulheres brancas transplante	0.10	$2.32e^{-17}$
Homens negros Oasis / homens brancos Oasis	0.04	0.003
Homens negros câncer / homens brancos câncer	0.14	$1.69e^{-34}$
Homens negros coração / homens brancos coração	0.21	$3.90e^{-71}$
Homens negros diabetes / homens brancos diabetes	0.11	$2.25e^{-18}$
Homens negros transplante / homens brancos transplante	0.10	$2.32e^{-17}$